nature methods

Article

https://doi.org/10.1038/s41592-024-02437-w

# De novo protein design with a denoising diffusion network independent of pretrained structure prediction models

Received: 28 November 2023

Accepted: 30 August 2024

Published online: 09 October 2024

Check for updates

Yufeng Liu[1,2,7], Sheng Wang[1,2,7], Jixin Dong[1,2], Linghui Chen[3], Xinyu Wang[1,2], Lei Wang[2], Fudong Li[2,4], Chenchen Wang[2], Jiahai Zhang[2,4], Yuzhu Wang[2], Si Wei[5], Quan Chen[1,2,3,4] ✉ & Haiyan Liu[2,3,4,6] ✉

The recent success of RFdiffusion, a method for protein structure design with a denoising diffusion probabilistic model, has relied on fine-tuning the RoseTTAFold structure prediction network for protein backbone denoising. Here, we introduce SCUBA-diffusion (SCUBA-D), a protein backbone denoising diffusion probabilistic model freshly trained by considering co-diffusion of sequence representation to enhance model regularization and adversarial losses to minimize data-out-of-distribution errors. While matching the performance of the pretrained RoseTTAFold-based RFdiffusion in generating experimentally realizable protein structures, SCUBA-D readily generates protein structures with not-yet-observed overall folds that are different from those predictable with RoseTTAFold. The accuracy of SCUBA-D was confirmed by the X-ray structures of 16 designed proteins and a protein complex, and by experiments validating designed heme-binding proteins and Ras-binding proteins. Our work shows that deep generative models of images or texts can be fruitfully extended to complex physical objects like protein structures by addressing outstanding issues such as the data-out-of-distribution errors.

A major problem to be solved in de novo protein design, which seeks to generate artificial proteins tailored to specific functions[1–5], is to generate protein structures that are designable/physically plausible, that is, that can be autonomously adopted by certain amino acid sequences[6–9]. This problem was addressed by using understandings learned from known protein structures[9–13]. The most recent progress is RoseTTAFold diffusion or RFdiffusion[13], which adopts the denoising diffusion probabilistic models (DDPMs), a class of machine learning models that use learnt networks to denoise corrupted data[14–16]. While showing unparalleled performances in extensive experimental tests, RFdiffusion relied on differently fine-tuning the pretrained structure prediction network RoseTTAFold[17] on different tasks of denoising protein backbones.

It is valuable to develop freshly trained DDPMs to complement models like RFdiffusion: the independent network configurations and training of freshly trained DDPMs would allow them to avoid inheriting potential specific biases in existing structure prediction networks. However, freshly trained DDPMs of protein structures reported so far have met with difficulties in generating defect-free protein backbones

[1]Department of Rheumatology and Immunology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, Hefei National Research Center for Physical Sciences at the Microscale, Center for Advanced Interdisciplinary Science and Biomedicine of IHM, University of Science and Technology of China, Hefei, China. [2]MOE Key Laboratory for Membraneless Organelles and Cellular Dynamics, School of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China. [3]Oristruct Biotech Co. Ltd, Hefei, China. [4]Biomedical Sciences and Health Laboratory of Anhui Province, Anhui Basic Discipline Research Center of Artificial Intelligence Biotechnology and Synthetic Biology, University of Science and Technology of China, Hefei, China. [5]iFLYTEK Research, Hefei, China. [6]School of Biomedical Engineering, Suzhou Institute for Advanced Research, University of Science and Technology of China, Hefei, China. [7]These authors contributed equally: Yufeng Liu, Sheng Wang.
✉e-mail: chenquan@ustc.edu.cn; hyliu@ustc.edu.cn

that can be realized with existing sequence design methods and confirmed by experimentally determined structures[12,18–22]. We reasoned that these difficulties could be attributed to the fact that DDPMs were usually trained by considering only the objective of maximizing the probability of recovering the training data (that is, the data recovery objective). While this objective is sensitive to errors of failing to generate certain observed data (that is, the model is required to be able to generate diverse structures), it is insensitive to errors of generating data of the real data distribution (that is, the model is not directly punished for generating out-of-distribution data)[23,24] (Fig. 1a). It is the data-out-of-distribution errors that cause faulted or unrealizable backbones. An efficient approach to reduce such errors is to consider the additional objective of minimizing adversarial losses as in the so-called generative adversarial networks (GANs)[25], in which discriminator networks are jointly trained with generator networks and the probability for generating data distinguishable from the training data is minimized.

Here, through training with combined objectives of data recovery and of minimizing adversarial losses, we developed a freshly trained DDPM that can generate diverse protein backbones with accuracies confirmed by experiments. We named our model SCUBA (sidechain unknown backbone arrangement)-diffusion or SCUBA-D as it produced designable backbones without predetermined amino acid sequences. We demonstrated that SCUBA-D can perform a variety of protein design tasks (Fig. 1b), including generating designable backbones from random noise (unconditional generation), generating designable backbones around user-sketched, undesignable initial backbones (generation based on sketch input) and generating backbones to scaffold predefined motifs with functions to bind small molecules or to bind other proteins (motif scaffolding). We verified SCUBA-D for these tasks by using either the ABACUS-R[26] or the ProteinMPNN[27] programs to select amino acid sequences for the generated backbones and experimentally characterizing an extensive set of designed proteins, including obtaining X-ray crystal structures of 16 de novo proteins and a protein complex and verifying the ligand binding function of a number of designed heme-binding proteins and several designed proteins that bind the human protein Ras (regulators of signal transduction; Fig. 1c).

## Results

### An overview of the SCUBA-D model

The network architecture of SCUBA-D is overviewed in Fig. 1d and described in details in the Methods. Briefly, two modules are used to carry out denoising from an initial backbone $\boldsymbol{x}_{\text{init}}$ to a final backbone $\tilde{\boldsymbol{x}}_0$ ($t = 0$). The first module performs one-step denoising to generate a low-resolution model $\boldsymbol{\mu}_{\boldsymbol{x}}$, which is used as nonzero priors and further denoised in multiple, successive diffusion steps in the second module[28,29]. Inside the modules, the backbone structure is represented by a pair representation. A single representation of amino acid sequence is used to guide the training of the pair representation. Details of initializing and updating the pair and the single representations are presented in the Methods. We note that structural information contained in the pair representation is used to update the single representation. This flow of information allows the use of the single representation training loss as an extra model regularization term to incorporate knowledge about natural amino acid sequences during training.

SCUBA-D was trained by using natural protein structures corrupted with various levels of noises as initial backbones. Details of the training data and training losses are described in the Methods. Among the major training losses shown in Fig. 1d, the 'frame aligned point error' (FAPE) loss[30] measures structural deviation, and the single representation loss measures the sequence deviation in a pretrained evolutionary scale modeling (ESM)[31] representation space belonging to data recovery losses, while the adversarial losses for reducing the data-out-of-distribution errors were provided by two discriminator subnetworks, one processing the local backbone conformation and the other processing inter-residue packing.

### Impacts of sequence representations and adversarial losses

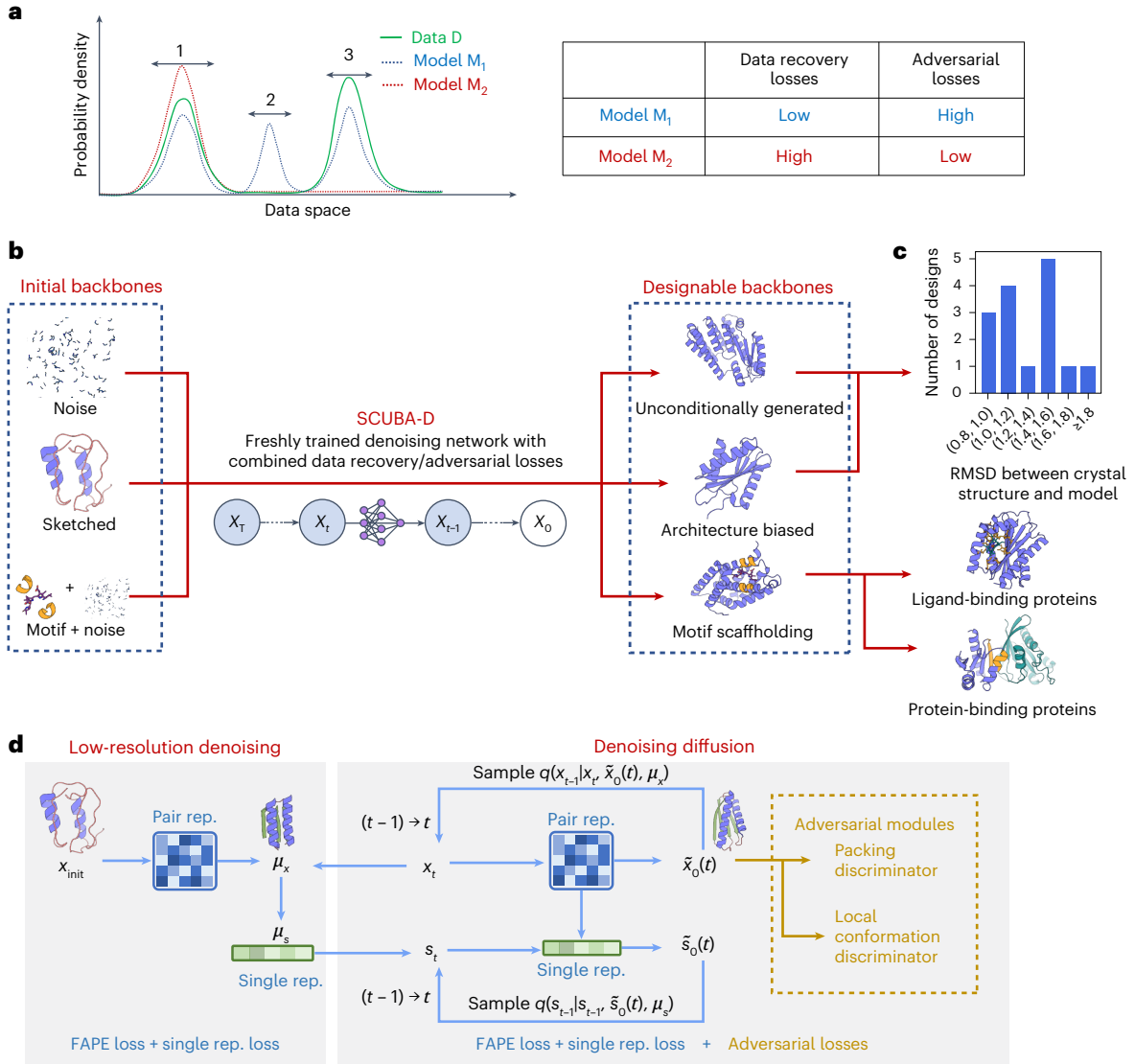
To examine these impacts, we compared four variant models. Three of the variant models were trained without any adversarial loss and with different learning targets for the single representation: the 'no ESM' model trained without using the ESM vector encoding native sequences as the learning target; the 'compressed ESM' model using a compressed version of the ESM vector as the learning target (Methods); and the 'full ESM' model using the original ESM vector in full dimensions as the learning target. The fourth variant model 'full ESM with GAN' was trained with the 'full ESM' learning target and with adversarial losses. To facilitate comparisons, the models were trained on the same data (up to the same number of 70,000 updating steps). The variant models were applied to 'denoise' 25 natural input backbones covering three fold classes (all-α, all-β and mixed αβ) whose CATH (class, architecture, topology and homologous superfamily)[32] topology types did not occur in the training data (Methods). For each input backbone, 3 'denoised' backbones were obtained using each variant model. The one-step denoising by all the variant models retained the natural backbones with negligible changes (root mean square deviation or RMSD < 0.7 Å). Subsequently the DDPM denoising modules produced backbones varied around these nonzero priors (here the natural backbones).

The output backbones were evaluated using three approaches, and the corresponding results are summarized in Extended Data Fig. 1a–c. First, deviations of the 'denoised' backbones from the corresponding input backbones were measured by the template modeling scores (TM scores)[33] and RMSDs of atomic positions (Extended Data Fig. 1a). We note that the possible values of TM scores are between 0 to 1 with a higher value representing closer structure resemblance. A TM score of 0.5 or lower indicates two structures of dissimilar overall folds[34]. Second, the deviations of the 'denoised' backbones from those predicted by AlphaFold2 (AF2)[30] on amino acid sequences selected for these backbones with the ABACUS-R program were measured as the self-consistent TM scores (scTM scores) and RMSDs (scRMSDs; Extended Data Fig. 1b) to judge the designability of the backbones produced by the models. Third, designability was also assessed using the ABACUS-R logits score (Extended Data Fig. 1c), which measures the compatibility between a backbone and the amino acid sequence optimally chosen (by ABACUS-R) for that backbone[26].

The results in Extended Data Fig. 1a–c show large improvements of 'full ESM' over 'compressed ESM' and 'no ESM' both in retaining the input natural backbones and in producing designable 'denoised' backbones. This indicated that positive model regularization took place by using 'full ESM' as the learning target of the single representation. Nevertheless, 'full ESM' still produced 'denoised' backbones of lower ABACUS-R logit scores than the natural backbones (Extended Data Fig. 1c), suggesting defects in the 'denoised' structures. The 'full ESM with GAN' model performed the best among the four variant models and led to comparable ABACUS-R logit scores as the natural backbones. Extended Data Fig. 1d shows a specific example of a 'denoised' backbone from 'full ESM with GAN' producing much smaller scRMSDs than the 'denoised' backbone from 'full ESM' (0.96 Å versus 5.45 Å) despite that the two 'denoised' backbones were similar (RMSD 1.39 Å). Based on the evaluations of the variant models, we trained the 'full ESM with GAN' model until the total loss stopped decreasing. The resulting network parameters were used in subsequent computational and experimental tests.

We examined the possibility of obtaining amino acid sequences directly by applying the ESM residue type classifier network[31] on the single ESM representations produced by SCUBA-D. The resulting sequences (noted as the pESM sequences) bore considerable similarity to sequences designed with ProteinMPNN[27] on the same backbone structures (for the aforementioned 25 natural backbones, the averaged ratio of physicochemically similar residue types between the two sets of sequences was 51%, with the similarity determined by dividing the 20 residue types into five similarity groups including

https://doi.org/10.1038/s41592-024-02437-w




| | Data recovery losses | Adversarial losses |
|---|---|---|
| Model $M_1$ | Low | High |
| Model $M_2$ | High | Low |

**Fig. 1 | SCUBA-D uses a denoising diffusion network trained with adversarial losses to generate designable protein backbone structures. a**, Three hypothetical probability distributions illustrating out-of-distribution errors and data recovery errors. The true data distribution (green solid line) comprises two density peaks, one in region 1 and the other in region 3. The model $M_1$ (blue dotted line) contains a spurious peak in region 2, representing the out-of-distribution errors. The model $M_2$ (red dotted line) misses the density peak in region 3 of the true distribution, indicating the data recovery errors. **b**, SCUBA-D uses a freshly trained denoising diffusion network to generate designable backbones from variously noised initial backbone structures, including initial structures composed of random noises ('unconditional generation'), sketched initial structures of certain approximately defined overall folds, or structures composed of well-defined parts combined with random noises (for 'motif scaffolding'). **c**, The crystal structures of multiple de novo proteins generated using SCUBA-D either unconditionally or with biases toward certain architectures were solved with resolutions of 1.3–2.6 Å. The numbers of structures with backbone RMSD values between the generated and the experimentally determined crystal structures in different ranges are shown. The use of SCUBA-D for motif scaffolding was assessed by generating and experimentally characterizing proteins that bind to a small-molecule ligand and proteins that bind to a target protein. **d**, SCUBA-D generates designable protein backbones based on two sequentially connected network modules to denoise initial protein backbone structures $x_{init}$. The first module performs low-resolution denoising on $x_{init}$ in one step. The second module uses the output model $\mu_x$ from the first module as nonzero prior data to perform further denoising diffusion over a number of steps. The training losses for the low-resolution denoising module include the structural FAPE loss and the sequence embedding single representation loss; the training losses for the denoising diffusion module additionally include adversarial losses contributed by two subnetwork modules distinguishing between natural and generated backbones: one based on spatial inter-residues geometry (packing), and the other based on sequential internal coordinates for continuous peptides (local conformation).

GAVLI, FYW, CM, ST, KRH, DENQ and P). However, the pESM sequences recovered only 20% of the corresponding native amino acid residue types exactly, while the same ratio of exactly recovered residue types by ProteinMPNN was 61%. Moreover, Extended Data Fig. 1e shows that the pESM sequences are of much lower ABAUCS-R logit scores than the native sequences or the ProteinMPNN-designed sequences, indicating that the pESM sequences are much less compatible with the corresponding backbones than the natural sequences or the sequences posteriorly designed based on given backbones.

## Generating protein backbones without conditioning

We compared SCUBA-D with five other DDPM-based methods (RFdiffusion[13], Chroma[21], FrameDiff[22], Genie[35] and ProteinSGM[12]) by unconditionally generating backbones of lengths from 100 to 400 residues with different methods and examining the resulting backbones' scRMSDs (calculated with ESMfold[31] on ProteinMPNN-designed sequences), numbers of clashed residues (evaluated by checking all interatomic distances against a sum of atomic radii plus a tolerance factor of 1.5 Å), and highest TM scores to Protein Data Bank (PDB)

structures. The summarized results are reported in Fig. 2a–c and Extended Data Fig. 2a.

For the groups of 100-residue backbones (each group comprised 100 backbones unconditionally generated with a given method), the mean scRMSDs of the RFdiffusion (0.72 Å) and the SCUBA-D (0.78 Å) backbones are substantially lower than the mean scRMSDs of the backbones generated with the other methods (from 2.36 to 3.72 Å). For the groups of backbones of 200–400 residues (each group comprised 300 backbones of 200, 300 or 400 residues), the mean scRMSDs of the RFdiffusion (3.28 Å) and the SCUBA-D (3.46 Å) backbones are also substantially lower than those of the Chroma (8.15 Å) and the FrameDiff (9.32 Å) backbones (the other two methods, Genie and ProteinSGM, cannot generate backbones of these lengths). The numbers of backbones with scRMSD below 2.5 Å listed in Extended Data Fig. 2a indicate similar relative performances of the different methods.

Notably, the results in Fig. 2c and Extended Data Fig. 2a suggest that SCUBA-D and Chroma produce substantially higher fractions of novel structures (highest TM score to PDB below 0.5) for proteins of 100 residues than RFdiffusion and FrameDiff. Extended Data Fig. 2b shows two examples of SCUBA-D-generated 100-residue backbones with the highest TM scores below 0.5 to either the database of PDB structures or the database of AF2-predicted structures[30], which indicate little overall structure similarity between the two generated backbones and the respective matched structures in the databases. For backbones of 200–400 residues, all four methods produce high fractions (more than one-third) of novel structures (Fig. 2c and Extended Data Fig. 2a). This should be because the conformational space for longer proteins is much more extensive, leaving many physically possible regions in this space uncovered by the limited number of structures in the PDB database. Moreover, a model biased toward generating small structures of relatively low novelty can still yield large structures of substantial overall novelty by combining various low-novelty small units.

One advantage of freshly trained DDPMs is that they can avoid potential biasing errors of a pretrained structure prediction network. To further examine this, we generated 500 backbones of 100 residues separately with SCUBA-D and RFdiffusion, and visualized the distribution of these backbones together on a two-dimensional plane by using *t*-distributed stochastic neighbor embedding (*t*-SNE)[36]. The resulting plot (Fig. 2d) shows that while the backbones populate an overall spherically shaped region, the RFdiffusion backbones densely populate only the upper-right part and are rarely found in the lower-left part. The SCUBA-D backbones are distributed over the entire sphere, covering both the lower-left part rarely sampled by RFdiffusion and the upper-right part populated by the RFdiffusion backbones. Thus, by using an orthogonally trained DDPM, SCUBA-D generates backbones that are complementarily distributed in the space of designable backbone structures in comparison with RFdiffusion.

Among the 500 SCUBA-D-generated backbones, we identified 80 backbones of both high structural novelty (highest TM scores to PDB structures below 0.5) and high predicted designability (scRMSDs obtained with ESMfold on ProteinMPNN sequences below 2.0 Å). We investigated the ability of the RoseTTAFold network to predict the structures of these backbones. The scatterplot in Fig. 2e indicates that for 12 of these backbones, RoseTTAFold2 was not able to predict their structures from the ProteinMPNN-designed sequences (the corresponding scRMSDs are above 6 Å). Visual inspections of these structures did not reveal any features that could potentially impair designability (see Fig. 2e for two example structures and Extended Data Fig. 2c for the remaining ten structures). These structures serve as example cases of novel backbones different from those predictable with the RoseTTAFold network.

Finally, we obtained a set of unconditionally generated SCUBA-D backbones of chain lengths evenly covering 100 to 490 residues at ten-residue intervals and evaluated their scRMSDs and highest TM scores to PDB (Supplementary Fig. 1a,b), which showed similar distributions as the corresponding results in Fig. 2a,c, respectively. From this set, we selected 16 backbones of around 200 residues in length, of scRMSDs < 2.0 Å, and of diverse backbone topologies (through visual inspections), and conducted experimental characterizations of amino acid sequences designed (with ProteinMPNN) on these backbones. We successfully purified 12 designed proteins in soluble form and confirmed their monomeric state by size-exclusion chromatography (SEC; see example results in Extended Data Fig. 3). We obtained protein crystals for 7 proteins, and solved 6 X-ray structures (2 all-α and 4 mixed αβ) with resolutions from 1.3 Å to 2.1 Å (Fig. 2f,g; see Supplementary Table 1 for summary of X-ray data). The crystal structures agree well with the corresponding designed structures with backbone RMSDs from 0.96 Å to 1.73 Å. The RMSDs for 38 of the 41 loop regions in these proteins are also below 2.0 Å (Extended Data Fig. 4a).

## Biasing the secondary structure distributions

In denoising diffusion without conditioning, helices may form more easily than β-sheets because a helix comprises a contiguous segment of residues in the primary sequence, while a β-sheet is composed of multiple separated segments. A previous analysis has suggested the favoring of helices in unconditional backbone generation by DDPMs[22]. We also noticed that no experimentally confirmed all-β backbones generated with DDPMs have been reported. We looked into this issue by examining first the mutual TM scores (Extended Data Fig. 5a) and then the secondary structure compositions of the unconditionally generated backbones (Extended Data Fig. 5b). While the TM score distribution suggested that the backbones were not concentrated with particular overall folds, the secondary structure composition distributions indicated that the backbones were enriched in helices and in lack of β-sheets in comparison with natural proteins.

To generate backbones with more β-sheets, we implemented a simple approach to bias the secondary structure states along the backbones by taking advantage of the nonzero priors in SCUBA-D (Methods). Extended Data Fig. 5c shows that the majority of backbones generated with this approach could reproduce the intended secondary structure states at specific positions with rates > 70%, and these backbones were of similar scRMSD distributions as the backbones generated without conditioning. For every intended secondary structure distribution, backbones with small scRMSD values (<2 Å) and high secondary structure state recovery rate (>0.7) could be generated, although the all-α distributions were targeted with higher success rates than the β-containing distributions (Extended Data Fig. 5d).

We also experimentally characterized the amino acid sequences obtained by applying ProteinMPNN on 17 backbones (4 of all-α fold, 5 of mixed αβ fold and 9 of all-β fold) generated with biased secondary structure distributions. Among these proteins, 12 were soluble (4 of all-α fold, 3 of mixed αβ fold and 5 of all-β fold). We were able to obtain 3 protein crystals but could not obtain X-ray diffraction data suitable for structure determination. Considering that failures in determining high-resolution structures do not necessarily mean that the designed proteins are not well folded, we measured the nuclear magnetic resonance (NMR) $^{15}$N-$^{1}$H heteronuclear singular quantum correlation (HSQC) and circular dichroism (CD) spectra of DB6, a protein designed to be of an all-β structure. The results (Fig. 2h) indicate that DB6 is indeed well folded with its dominant secondary structure type being the β-sheet.

## Generating protein structures based on sketched inputs

Sketched input structures can be used by SCUBA-D to enable more focused explorations of the designable backbone space. We evaluated SCUBA-D for such tasks by considering sketched backbones adopting a given overall architecture specified by the types, approximate sizes and coarsely defined three-dimensional arrangements of secondary structure elements. First, we considered the architectures adopted by the aforementioned 25 natural proteins. Fig. 3a shows that for 16

https://doi.org/10.1038/s41592-024-02437-w

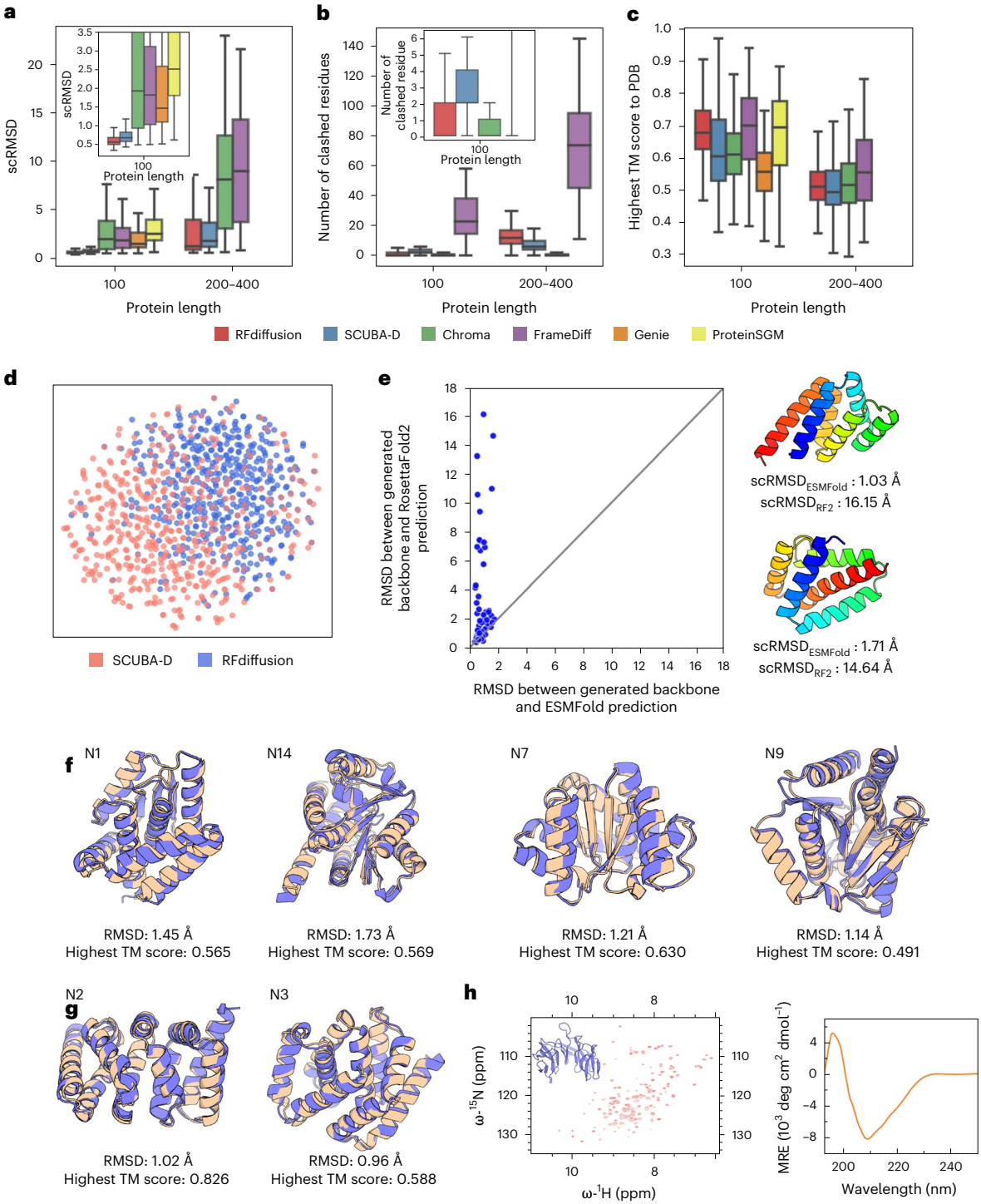



**Fig. 2 | Structure generation without condition or with biased secondary structure distributions. a**–**c**, Comparisons between different DDPM models. The box plots show median, interquartile range, and minimum and maximum values excluding outliers (>1.5 times the interquartile range beyond the box) for sample sizes of 100 backbones for a box representing the chain length of 100 residues or 300 backbones for a box representing the chain length of 200–400 residues (100 backbones for each of the chain lengths of 200, 300 and 400 residues). Results of different methods are colored differently, as indicated below the plots: scRMSDs (**a**), numbers of clashed residues per backbone (**b**) and highest TM scores to PDB structures (**c**). **d**, A two-dimensional visualization of 100-residue backbones unconditionally generated by SCUBA-D and by RFdiffusion (500 backbones with each method) obtained with *t*-SNE, with the inverses of the mutual TM scores treated as 'distances'. **e**, The scRMSDs of 80 SCUBA-D-generated backbones obtained by using RosettaFold2 for structure prediction compared with the scRMSDs obtained by using ESMFold for structure prediction. These backbones are of the highest TM scores to PDB structures below 0.5 and of ESMfold prediction-based scRMSDs below 2.0 Å. The structures of two example backbones with RosettaFold2 prediction-based scRMSDs above 14.0 Å are shown on the right. Both the ESMFold prediction-based scRMSDs and the RosettaFold2 prediction-based scRMSDs are indicated. **f**, Crystal structures (gold) of four designed mixed αβ proteins superimposed with the corresponding backbones generated by SCUBA-D (blue) without condition. For each designed protein, the protein ID, the RMSD between the crystal structure and the generated backbone, and the highest TM score of the generated backbone to existing structures in the PDB are indicated. The amino acid sequences of the proteins were designed with ProteinMPNN. **g**, The same as **f**, but for two designed all-α proteins. **h**, The NMR $^{15}$N-$^{1}$H HSQC spectrum (left) and CD spectrum (right) of an all-β protein generated by SCUBA-D with a biased distribution of the secondary structure states along the peptide chain. The secondary structure contents estimated from the CD spectrum are 8.6% helix, 66.6% β strand and 6.3% turn.

https://doi.org/10.1038/s41592-024-02437-w

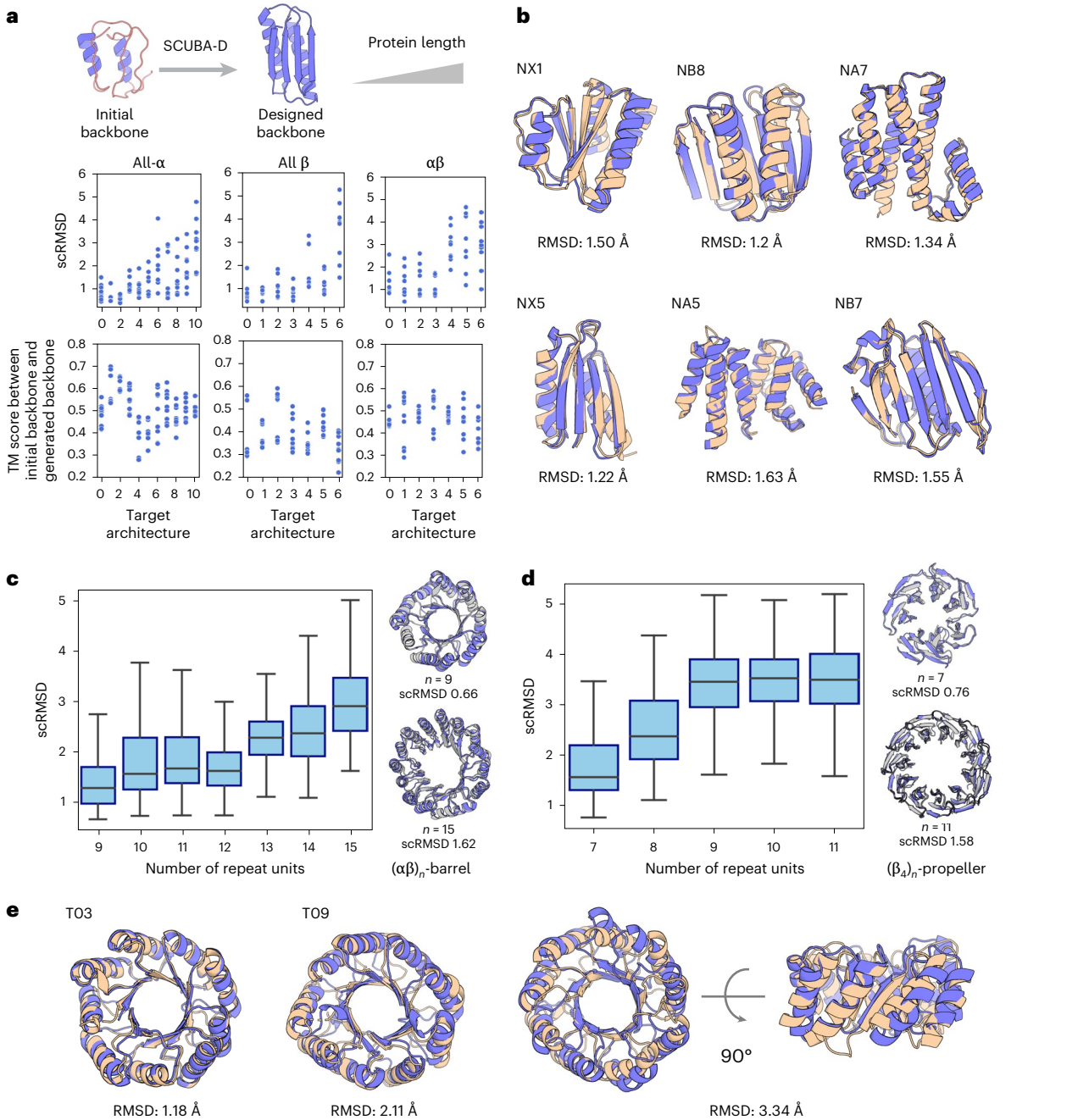



**Fig. 3 | Generating protein structures with sketched inputs. a**, The scRMSDs and the TM scores to the corresponding initial structures for backbones designed by SCUBA-D using initial structures 'sketched' according to the architectures of 25 natural proteins. For each architecture, nine backbones were generated and evaluated, one data point in the plots corresponding to one designed backbone. The results for architectures of the three different classes of secondary structure compositions (all-α, all-β and mixed αβ) are displayed in different plots. Within each plot, results for the same architecture are numbered the same and displayed in the same column, and the architectures are arranged from left to right in the ascending order of the chain lengths. **b**, Crystal structures (gold) of six designed proteins superimposed with the corresponding backbones (blue) generated by SCUBA-D based on initial backbones 'sketched' according to given architectures. The protein IDs and the RMSDs of the superimpositions are indicated. **c**, Distributions of scRMSDs of the $(\alpha\beta)_n$-barrel backbones generated by SCUBA-D. For each value of the number of repeat units $n$, four initial backbones were considered. The analysis was performed on 32 sequences generated by considering 8 independent SCUBA-D runs with each initial backbone. The box plots show median, interquartile range, and minimum and maximum values excluding outliers (>1.5 times the interquartile range beyond the box). Two example backbones ($n = 9$ and $n = 15$) superimposed with AF2-predicted structures for ProteinMPNN-designed amino acid sequences are shown next to the box plot. The corresponding scRMSDs are indicated. **d**, The same as **c**, but for the $(\beta_4)_n$-propeller backbones with $n$ ranging from 7 to 11. Two example backbones ($n = 7$ and $n = 11$) superimposed with AF2-predicted structures for ProteinMPNN-designed amino acid sequences are shown next to the box plot. **e**, Left and middle, crystal structures (gold) of two designed $(\alpha\beta)_9$-barrel proteins superimposed with the corresponding backbones (blue) generated by SCUBA-D; the protein IDs and RMSDs of the superimpositions are indicated. Right, superimposition of the crystal structures of the two designed proteins (T09 in blue, and T03 in yellow); the RMSD of the superimposition is indicated.

https://doi.org/10.1038/s41592-024-02437-w

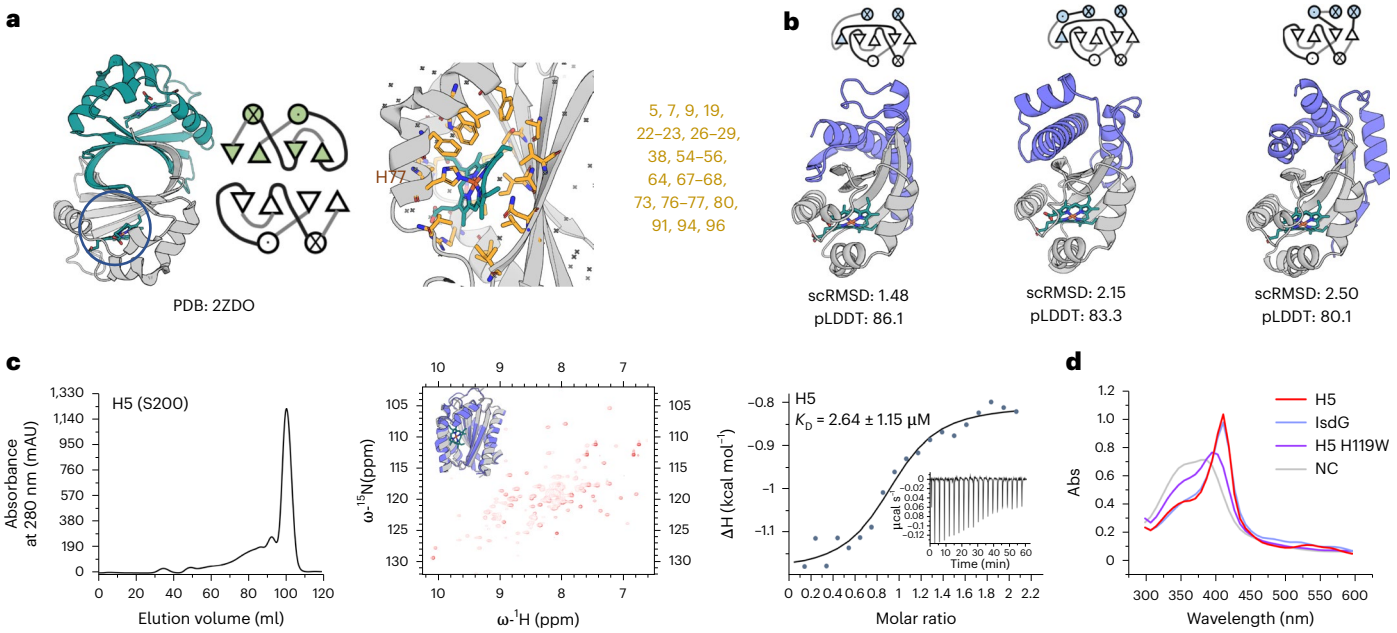



**Fig. 4 | SCUBA-D for designing small-molecule-binding proteins. a**, Left, the dimeric structure of the natural heme-binding protein IsdG (PDB 2ZDO); the two monomers are colored differently; the blue circle indicates the location of the heme-binding pocket in one monomer; the topology diagram of the dimer is shown next to the structure. Right, an enlarged view of the heme-binding pocket, showing the heme molecule (green sticks with the iron center colored in pink) surrounded by residues (sticks with carbon atoms colored in yellow) whose geometries were extracted for scaffolding by SCUBA-D; the histidine (H77) in coordination with iron is indicated; the sequential numbers of residues forming the heme-binding pocket are listed next to the structure. **b**, The structures of three backbones (H5, H6 and H8) generated by SCUBA-D to scaffold the heme-binding pocket extracted from IsdG. The corresponding topology diagrams are drawn above the structures. The indicated scRMSDs and pLDDT scores are based on AF2 predictions on the experimentally tested amino acid sequences, which were designed with ABACUS-R. **c**, Experimental characterizations of the designed heme-binding protein H5. Left, SEC result. Middle, NMR $^{15}$N-$^{1}$H HSQC spectrum shown with a superimposition of the designed structures (blue) and the structure predicted with AF2 (gray). Right, ITC measurements on heme binding. **d**, UV-visible absorption spectra of the designed heme-binding protein H5, the natural protein IsDG, and the H119W mutant of H5. Abs, absorbance; mAU, milli-absorbance unit; NC, negative control.

of the 25 architectures, at least one designable backbone retaining the input folds (scRMSD < 2.0 Å and TM score to initial backbone > 0.5) were successfully generated by considering 9 sketched inputs for each architecture. For 8 of the 9 remaining architectures, backbones meeting the specified in silico criteria could also be obtained by increasing the number of sketched inputs to 60 (see example results in Extended Data Fig. 6a). The remaining one architecture for which none of the backbones generated from 60 initial backbones could meet the specified in silico criteria is shown in Extended Data Fig. 6b.

We further examined using the sketched input approach to design architectures of cyclically organized repeat units, which included a series of $(\alpha\beta)_n$-barrels with the repeat number $n$ from 9 to 15 and a series of $(\beta_4)_n$ propellers with the repeat number $n$ from 7 to 11 (See Methods and Supplementary Methods for details of constructing the sketched inputs). The scRMSD distributions are shown in Fig. 3c,d. Extended Data Fig. 7a,b shows that for every architecture of a given repeat number, at least one backbone of a scRMSD value below 2.0 Å was generated. We note that the $(\alpha\beta)_n$ architecture with $n > 8$ and the $(\beta_4)_{11}$ architectures are rarely observed in natural proteins. Moreover, backbones generated for 10 of the 12 architectures are of the highest TM scores of below 0.5 for the database of PDB structures and the database of AF2-predicted structures. Thus, the sketched input approach can also be used to generalize structures beyond natural proteins.

We conducted experimental characterization on 25 proteins designed for the backbones generated from sketched inputs (to examine the joint use of SCUBA-D with sequence design programs other than ProteinMPNN, the amino acid sequences used in this batch of experiments were designed with the ABACUS-R program). We successfully purified 18 designed proteins in soluble forms and confirmed their monomeric state by SEC (see example results in Extended Data Fig. 3). We obtained crystals of seven proteins and solved X-ray structures for six with resolutions from 1.7 Å to 2.2 Å (see Supplementary Table 2 for summary data of the X-ray structures). The RMSDs between the designed backbones and the solved structures range from 1.2 Å to 1.6 Å (Fig. 3b). The RMSDs for 60 of the 78 loop regions in these proteins are also below 2.0 Å (Extended Data Fig. 4b).

We also experimentally characterized 12 proteins with backbones generated for the $(\alpha\beta)_9$-barrel architecture and with amino acid sequences designed by ABACUS-R. Among them, 11 proteins were soluble, and 4 X-ray crystal structures were solved (see Supplementary Table 3 for summary data of the X-ray structures). Two crystals (T03 and T09) showed monomeric structures that could be closely aligned with the corresponding designed backbones (Fig. 3e). Local structural differences were present between T03 and T09 (Fig. 3e). The crystals for T01 and for T11 showed domain-swapped dimeric and trimeric structures, respectively (Extended Data Fig. 7c,d). Despite this, the individual domains in these crystal structures still present the designed $(\alpha\beta)_9$-barrel architecture. Additionally, static light scattering experiments indicated that T01 and T11 exist as monomers in solution (Extended Data Fig. 7c,d).

## Designing heme-binding proteins with SCUBA-D

An important application of de novo protein backbone design is to generate scaffolds that can precisely maintain the three-dimensional geometry of a group of residues predefined to form a functional site for specific interactions with other molecules[20,37]. While it has been shown that RFdiffusion was able to address this problem by tuning the pretrained structure prediction network specifically for scaffolding functional sites[13], successful scaffolding of small molecules of protein-binding sites with a freshly trained structure generation network has not been experimentally demonstrated.

https://doi.org/10.1038/s41592-024-02437-w

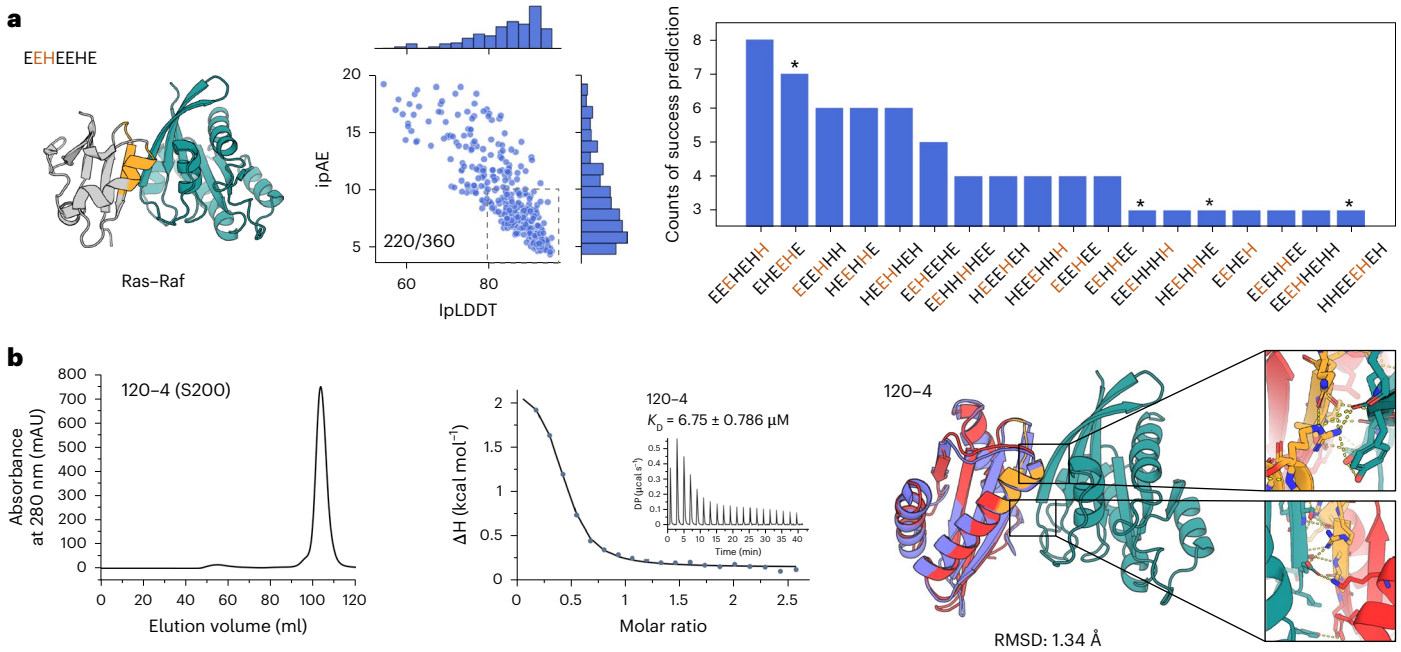



**Fig. 5 | SCUBA-D for designing protein-binding proteins. a**, Left, the structure of the complex between Ras (dark green) and the Ras-binding protein of Raf (gray; PDB 4G0N); the two Raf segments at the binding interface (yellow) were extracted for scaffolding by SCUBA-D; the sequential arrangement of the types of the secondary structure elements in the Raf domain is indicated as a string of the characters 'H' (for helix) and 'E' (for strand) above the structure, with the characters representing the two elements containing the scaffolded Ras-binding segments colored in orange. Middle, scatterplot and histograms of the lpLDDT and ipAE scores predicted by AF2-Multimer for the complexes between Ras and the designed Ras-binding proteins, whose backbones were generated with SCUBA-D and amino acid sequences were designed with ProteinMPNN. Right, histogram of the sequential arrangements of secondary structure elements of designed Ras-binding proteins with ipAE scores below 10 and lpLDDT scores above 80; the characters representing the two elements containing the scaffolded segments are colored in orange; the asterisks label secondary structure element arrangements for which at least one designed protein of each of these arrangements was verified to bind Ras by the DHFR-based protein complementarity analysis assay. **b**, The designed structure and the experimental characterizations of the Ras-binding protein of 120–4. From left to right, SEC experiment of the designed protein of 120–4; ITC for determining the binding with Ras; the superimposition of the designed structure of 120–4 (blue) and the crystal structure of the designed protein of 120–4 (red) in complex with Ras (green). The interface segments inherited from Raf are colored in yellow. Enlarged views of two interfaces of the crystal structure are shown next to the overall superimposition (with the inherited interface and the extended interface shown in the upper and lower panels, respectively). ΔH, change of enthalpy; DP, differential power.

We evaluated SCUBA-D for scaffolding small-molecule binding sites by generating backbones to hold single heme-binding sites extracted from a native iron-regulated surface determinant G protein (IsdG), whose structure in complex with heme has been solved in a dimeric state (PDB 2ZDO; Fig. 4a)[38]. Residues surrounding heme in one monomer of 2ZDO (pocket residues) were used as the motif to be scaffolded. More details of motif definition and backbone generation are described in Supplementary Methods. From a total of 90 sequences designed on 30 backbones with ABACUS-R, we identified 14 designed proteins predicted by AF2 with scRMSD values below 2.0 Å and ligand-predicted local distance difference test (pLDDT) values above 80 (Extended Data Fig. 8a). These designed proteins exhibited varied backbone architectures. Experimental characterizations were conducted on 12 designed proteins. Ultraviolet (UV)-visible spectrophotometry experiments[39] indicated that 5 of them could bind heme (Extended Data Fig. 8b). We measured the dissociation constants ($K_D$) for heme-binding isothermal titration calorimetry (ITC) for three of the proteins (Fig. 4b) showing comparable absorption peaks in the UV-visible spectra as the natural protein. SEC and NMR $^{15}$N-$^{1}$H HSQC spectra confirmed that these designed proteins form well-folded monomers in solution (Fig. 4c and Extended Data Fig. 8c,d). The measured $K_D$ values were 2.64 μM, 0.855 μM and 0.924 μM for H5, H6 and H8, respectively (Fig. 4c and Extended Data Fig. 8c,d), which could be compared with the $K_D$ values of 2.98 μM and 3.94 μM for the two heme-binding sites of the dimeric IsdG (Extended Data Fig. 8e). UV-visible spectra of mutants (H119W for H5, H129W for H6, and H131W for H8) confirmed the contribution of a histidine for coordination with the iron of heme in the designed proteins (Fig. 4d and Extended Data Fig. 8f).

## Designing Ras-binding proteins with SCUBA-D

To illustrate the design of protein binders, we used SCUBA-D to generate backbones to scaffold the binding sites presented by the Ras-binding domain of the protein Raf (rapidly accelerated fibrosarcoma). Two segments from Raf at the interface of the Ras–Raf complex[40] (PDB 4G0N; Fig. 5a) were used to compose initial input structures. The AF2-Multimer program[41] was used to filter 1,800 sequences designed by ProteinMPNN on 360 SCUBA-D-generated backbones. For 220 backbones, which were of diversely composed and ordered secondary structure elements, at least one designed sequence for each backbone was predicted with lpLDDT values > 80 and the interface predicted aligned error (ipAE) values < 10 (Fig. 5a; see Extended Data Fig. 9a for three example backbones and their predicted structures).

We experimentally assessed the binding of 30 designed proteins with Ras using the dihydrofolate reductase-based protein complementarity analysis assay[42], in which 14 designed proteins showed detectable binding, with 7 proteins producing results comparable to Raf (Extended Data Fig. 10a). Competitive protein complementarity analysis on 4 designed proteins showed that coexpressed Raf competitively inhibited the binding of the designed proteins with Ras (Extended Data Fig. 10b). These proteins are monomers in solution (see example SEC curves in Fig. 5b and Extended Data Fig. 3). To examine if the designed proteins were indeed well folded, we considered 90–2 as an example and measured its NMR $^{15}$N-$^{1}$H HSQC spectrum in solution, which

confirmed that the protein was well folded (Extended Data Fig. 9b). We were able to measure (using ITC) the $K_D$ values for Ras binding by three purified designed proteins, 90–2, 120–4 and 90–4, to be 6.70 μM, 6.75 μM and 13.0 μM, respectively (see Fig. 5b and Extended Data Fig. 9c), which indicated weaker binding than Raf ($K_D$ ~ 0.80 μM; Extended Data Fig. 9d).

The predicted complex structures (Extended Data Fig. 9a) of the designed proteins with Ras suggested that the interfaces in these complexes were more extended than in the Ras–Raf complex. Indeed, mutating an arginine residue (Arg89 in Raf) contained in the retained motif into leucine, while abolishing Ras binding in Raf and 90–4, only moderately weakened Ras binding in 90–2 and 120–4 (Extended Data Fig. 9c,d). To examine possible contributions of residues in the predicted extended binding interfaces in the latter two proteins, we investigated mutations p.Gly41Glu of 90–2 and p.Lys17Glu of 120–4. The p.Gly41Glu mutation of 90–2 weakened the Ras binding to a $K_D$ of 220 μM, while the p.Lys17Glu mutation of 120–4 abolished Ras binding. Finally, the structure of the complex between the designed protein 120–4 and Ras was confirmed by high-resolution crystal structure (Fig. 5b) with an overall RMSD of 1.34 Å (see Supplementary Table 4 for summary data of the X-ray structures). The RMSD for residues in the retained Raf motif was 0.38 Å. The RMSD of residues forming the extended interface was only 0.11 Å.

## Discussion

As structure prediction networks were trained by emphasizing the building of correct structural models for natural proteins based on their given sequences, not on generating (potentially novel) designable structures not constrained by particular sequences, potential biases could be introduced in the structures generated by models tuned from pretrained structure prediction networks. By using a freshly and orthogonally trained network for structure denoising, SCUBA-D can avoid inheriting such biases. We note that from only a limited number (500) of SCUBA-D backbones, a number of structures not observed in PDB and that cannot be correctly predicted by RoseTTAFold could be identified. Given the vastness of the protein structure space, the absolute number of designable backbones inaccessible to individual structure generation methods with biasing errors can be immense. The presence of these backbones suggests that even for proteins of small sizes, the known protein structures are still far from exhaustively covering all physically possible folds. Our results highlight the value of developing protein design models orthogonal to existing structure prediction networks.

Unlike other freshly trained DDPMs of protein structures that only considered the usual data recovery objectives, SCUBA-D has been trained with the additional objectives of minimizing adversarial losses. This has enabled SCUBA-D to generate backbones of much higher designability than other freshly trained DDPMs. To the best of our knowledge, experimentally solved structures of proteins designed with freshly trained DDPMs are so far limited to Chroma and SCUBA-D. The Chroma study considered more than 300 designed proteins for experimental characterization and reported the crystal structures of only two all-helical proteins[21], reflecting the difficulties of DDPMs trained with the usual data recovery objectives in generating protein structures confirmable with high-resolution experiments. In comparison, with the additional objective of minimizing adversarial losses for model training, SCUBA-D achieved substantially more extensive experimental successes than other freshly trained DDPMs. Moreover, the successful design of an all-β protein with SCUBA-D may be considered as an important step of protein backbone design with DDPMs. These results underscore the importance of considering the different error tolerance of physically constrained objects such as protein structures in comparison with nonphysically constrained objects such as images and texts. This insight can be utilized to expedite the extension of deep generative methods with demonstrated powers in generating in silico objects to the generation of objects that need to be physically plausible, including designable structures of nucleic acids and of protein–nucleic acid complexes.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-024-02437-w.

## References


1. Huang, P. S., Boyken, S. E. & Baker, D. The coming of age of protein design. *Nature* **537**, 320–327 (2016).
2. Kuhlman, B. et al. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
3. Polizzi, N. F. & DeGrado, W. F. A defined structural unit enables de novo design of small-molecule–binding proteins. *Science* **369**, 1227–1233 (2020).
4. Gainza, P. et al. De novo design of protein interactions with learned surface fingerprints. *Nature* **617**, 176–184 (2023).
5. Yeh, A. H.-W. et al. De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).
6. Li, H., Helling, R., Tang, C. & Wingreen, N. Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666–669 (1996).
7. Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA* **97**, 10383–10388 (2000).
8. Grigoryan, G. & DeGrado, W. F. Probing designability via a generalized model of helical bundle geometry. *J. Mol. Biol.* **405**, 1079–1100 (2011).
9. Huang, B. et al. A backbone-centred energy function of neural networks for protein design. *Nature* **602**, 523–528 (2022).
10. Anishchenko, I. et al. De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).
11. Eguchi, R. R., Choe, C. A. & Huang, P.-S. Ig-VAE: generative modeling of protein structure by direct 3D coordinate generation. *PLoS Comput. Biol.* **18**, e1010271 (2022).
12. Lee, J. S., Kim, J. & Kim, P. M. Score-based generative modeling for de novo protein design. *Nat. Comput. Sci.* **3**, 382–392 (2023).
13. Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
14. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
15. Chen, N. et al. Wavegrad: estimating gradients for waveform generation. In *Proc. International Conference on Learning Representations* (ICLR, 2021).
16. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with clip latents. Preprint at https://arXiv.org/abs/2204.06125 (2022).
17. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
18. Anand, N. & Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. Preprint at https://arXiv.org/abs/2205.15019 (2022).
19. Wu, K. E. et al. Protein structure generation via folding diffusion. *Nat. Commun.* **15**, 1059 (2022).
20. Trippe, B. L. et al. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *Proc. International Conference on Learning Representations* (ICLR, 2023).
21. Ingraham, J. et al. Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).

22. Yim, J. et al. SE(3) diffusion model with application to protein backbone generation. In *Proc. International Conference on Machine Learning* (ICML, 2023).
23. Zhao, H., Gallo, O., Frosio, I. & Kautz, J. Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging* **3**, 47–57 (2016).
24. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T. & Zelnik-Manor, L. The 2018 PIRM challenge on perceptual image super-resolution. In *Proc. the European Conference on Computer Vision (ECCV) Workshops* (eds Leal-Taixé, L. & Roth, S.) 334–355 (2019).
25. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
26. Liu, Y. et al. Rotamer-free protein sequence design based on deep learning and self-consistency. *Nat. Comput. Sci.* **2**, 451–462 (2022).
27. Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
28. Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., Kudinov, M. Grad-tts: a diffusion probabilistic model for text-to-speech. in *International Conference on Machine Learning*, 8599–8608 (PMLR, 2021)
29. Lee, S.-g. et al. PriorGrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. In *International Conference on Learning Representations* (ICLR, 2022).
30. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
31. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
32. Sillitoe, I. et al. CATH: increased structural coverage of functional space. *Nucleic Acids Res.* **49**, D266–D273 (2021).
33. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
34. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score=0.5? *Bioinformatics* **26**, 889–895 (2010).
35. Lin, Y. & AlQuraishi, M. Generating Novel, Designable, and Diverse Protein Structures by Equivariantly Diffusing Oriented Residue Clouds. In *Proc. International Conference on Machine Learning* (ICML, 2023).
36. van der Maaten, L. & Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn Res.* **9**, 2579–2605 (2008).
37. Wang, J. et al. Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
38. Lee, W. C., Reniere, M. L., Skaar, E. P. & Murphy, M. E. Ruffling of metalloporphyrins bound to IsdG and IsdI, two heme-degrading enzymes in Staphylococcus aureus. *J. Biol. Chem.* **283**, 30957–30963 (2008).
39. Skaar, E. P., Gaspar, A. H. & Schneewind, O. IsdG and IsdI, heme-degrading enzymes in the cytoplasm of Staphylococcus aureus. *J. Biol. Chem.* **279**, 436–443 (2004).
40. Fetics, S. K. et al. Allosteric effects of the oncogenic RasQ61L mutant on Raf-RBD. *Structure* **23**, 505–516 (2015).
41. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.04.463034 (2021).
42. Remy, I., Campbell-Valois, F. & Michnick, S. W. Detection of protein–protein interactions using a simple survival protein–fragment complementation assay based on the enzyme dihydrofolate reductase. *Nat. Protoc.* **2**, 2120–2125 (2007).

# Methods

## The DDPM used in SCUBA-D

In standard DDPMs, the process of adding noises to data (referred as the forward process) is defined as a Markovian diffusion process, while the denoising process is modeled with deep networks optimized to reproduce the conditional distributions in the reverse Markovian process. In the forward process, the initial distribution $q(\boldsymbol{x}_0)$ of the observed data $\boldsymbol{x}_0$ is gradually converted into a predefined distribution (usually a standard normal distribution) by adding noises through a Markov chain of $T$ discrete steps. The transition probability from step $t-1$ to step $t$ ($t \in [1, 2, \ldots, T]$) is usually defined to be a normal distribution, namely, $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right)$, in which $\{\beta_t, t \in [1, 2, \ldots, T]\}$ is called the noise schedule. The reverse (denoising) process generates a new data sample $\boldsymbol{x}_0$ from a noise sample $\boldsymbol{x}_T \sim p(x_T) = \mathcal{N}(0, \boldsymbol{I})$ with the learned transition probability $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \sigma_t^2 \boldsymbol{I})$, where $\theta$ is the model parameter and $\sigma_t = \frac{\beta_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$ with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \Pi_{s=1}^{t} \alpha_s$.

The standard DDPMs are trained by maximizing the evidence lower bound (ELBO) of the log likelihood of the training data, with the ELBO being the sum of the negative Kullback–Leibler (KL) divergences between the forward and reverse transition distributions of all the diffusion steps, according to equation (1)[13]:

$$\begin{aligned} &\log p(\boldsymbol{x}_0) \geq -D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\,|\,p(\boldsymbol{x}_T)) \\ &-\sum_{t=2}^{T} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)\,|\,p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \\ &-\log p(\boldsymbol{x}_0|\boldsymbol{x}_1). \end{aligned} \tag{1}$$

It can be shown that maximizing the above ELBO is equivalent to minimizing the following denoising loss, according to equation (2):

$$\mathcal{L}_\theta = E_{t,\boldsymbol{x}_0}\left[\frac{\beta_t^2 \bar{\alpha}_{t-1}}{2\sigma_t^2(1-\bar{\alpha}_t)^2}||\boldsymbol{x}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{x}_0||^2\right], \tag{2}$$

where $\boldsymbol{x}_\theta(\boldsymbol{x}_t, t)$ is the predicted denoised data (noted as $\tilde{\boldsymbol{x}}_0(t)$ below) based on $\boldsymbol{x}_t$. Here the noised data $\boldsymbol{x}_t$ is generated by scaling and adding noises to the observed data $\boldsymbol{x}_0$, according to equation (3):

$$x_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{3}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The denoising version of the DDPM loss in equation (2) simplifies the learning process into learning a denoising model $\boldsymbol{x}_\theta(\boldsymbol{x}_t, t)$ that can generate the predicted data $\tilde{\boldsymbol{x}}_0(t)$ to best reproduce the true data $\boldsymbol{x}_0$.

With a denoising model trained with the above scheme, new samples or synthesized data can be generated by iteratively applying the following two steps, starting from sampling the random distribution $p(\boldsymbol{x}_T)$,

1. predict the denoised sample for step $t$ as $\tilde{\boldsymbol{x}}_0(t) = \boldsymbol{x}_\theta(\boldsymbol{x}_t, t)$;
2. predict the diffused sample at step $t-1$ by adding noises to the denoised sample for step $t$, namely, $\boldsymbol{x}_{t-1} \sim q(\boldsymbol{x}_{t-1}|\tilde{\boldsymbol{x}}_0(t), \boldsymbol{x}_t) = \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(\beta_t)}{1-\bar{\alpha}_t}\tilde{\boldsymbol{x}}_0(t) + \tilde{\beta}_t\boldsymbol{\epsilon}$.

In this study, we adopted a recently proposed variant of the above standard DDPM model, the PriorDDPM or DDPM with nonzero priors[28,29]. In PriorDDPM, the end point of the forward diffusion process is defined to follow a Gaussian distribution with nonzero means, that is, $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I})$. This is achieved by modifying equation (3), which gives equation (4) as follows:

$$\boldsymbol{x}_t = \boldsymbol{\mu} + \sqrt{1-\beta_t}(\boldsymbol{x}_{t-1} - \boldsymbol{\mu}) + \sqrt{\beta_t}\boldsymbol{\epsilon} = \boldsymbol{\mu} + \sqrt{\bar{\alpha}_t}(\boldsymbol{x}_0 - \boldsymbol{\mu}) + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}. \tag{4}$$

Similarly to standard DDPM, the PriorDDPM can be trained with a denoising loss as defined in ref. 2. During training and inference, the denoising network can now be guided by the prior, namely, $\tilde{\boldsymbol{x}}_0(t) = \boldsymbol{x}_\theta(\boldsymbol{x}_t, \boldsymbol{\mu}, t)$.

In SCUBA-D, the nonzero prior $\boldsymbol{\mu}$ in equation (4) is generated by the low-resolution denoising module (Fig. 1d). This module works like a denoising autoencoder that generates a low-resolution backbone from an input initial structure in a single step.

## The representation of backbones and the denoising blocks

In SCUBA-D, the backbone structures noted as $\boldsymbol{x}$ above are actually a collection of the positions and orientations of individual residues. For each residue, its position and orientation are specified by a rigid body transformation $\hat{T} = (\boldsymbol{T}, \boldsymbol{R})$, in which the three-dimensional vector $\boldsymbol{T}$ denotes translation of the Cα atom coordinates relative to the origin of a global coordinate frame, while $\boldsymbol{R}$ denotes the rotation of a local coordinate frame (that is, a coordinate frame fixed on the residue) relative to a fixed reference frame in the global coordinate system, which is also represented by a three-dimensional vector of three angular values. The components of $\boldsymbol{R}$ are related to the quaternion representation of the rotation, $\boldsymbol{q} = [q_1, q_2, q_3, q_4]$, with the relations as shown in equation (5):

$$\begin{bmatrix} \theta \\ \phi \\ \psi \end{bmatrix} = \begin{bmatrix} \arccos & (q_1) \\ \arctan & (q_2|q_3) \\ \arccos & \left(\frac{q_4}{\sqrt{1-q_1^2}}\right) \end{bmatrix}, \tag{5}$$

The structure denoising blocks in SCUBA-D, that is, the low-resolution denoising module and the model $\boldsymbol{x}_\theta(\boldsymbol{x}_t, \boldsymbol{\mu}, t)$ described above, were developed based on the ideas of AF2 (ref. 30) for transforming protein structures with deep networks that are equivariant to translations and rotations. Specifically, input structures containing noises represented by the above set of residue-wise $(\boldsymbol{T}, \boldsymbol{R})$ features are first transformed into residue pairwise features of $N_{\mathrm{res}} \times N_{\mathrm{res}}$ units ($N_{\mathrm{res}}$ denotes the number of residues) corresponding to the relative translations and rotations between each pair of residues. The resulting pair representation is updated using triangle attentions, which is followed by Invariant Point Attention[30] for retrieving the residue-wise $(\boldsymbol{T}, \boldsymbol{R})$ features of the denoised output structure. When these equivariant denoising blocks are used together with invariant training losses, the resulting overall network is equivariant to translation and rotation.

In SCUBA-D, the denoising diffusion of the structure is assisted with a co-diffused single representation of amino acid sequence (noted as $\boldsymbol{s}$ below) with $N_{\mathrm{res}}$ units. The nonzero priors of the single representation $\boldsymbol{\mu}_{\boldsymbol{s}}$ is predicted from the denoised prior structure by an extra decoder (implemented as a Geometry Vector Perceptron subnetwork[43] in the low-resolution denoising module). In the DDPM module, the single representation is denoised in the same way as the structure, namely, according to equations (6) and (7):

$$\boldsymbol{x}_t = \boldsymbol{\mu}_{\boldsymbol{x}} + \sqrt{\bar{\alpha}_t}(\boldsymbol{x}_0 - \boldsymbol{\mu}_{\boldsymbol{x}}) + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{x} \text{ and} \tag{6}$$

$$\boldsymbol{s}_t = \boldsymbol{\mu}_{\boldsymbol{s}} + \sqrt{\bar{\alpha}_t}(\boldsymbol{s}_0 - \boldsymbol{\mu}_{\boldsymbol{s}}) + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{s}, \tag{7}$$

in which $\boldsymbol{\mu}_{\boldsymbol{x}}$ and $\boldsymbol{\mu}_{\boldsymbol{s}}$ denote the priors of the structure and of the sequence representation, respectively. The denoising subnetwork for the single representation can be noted according to equation (8):

$$\tilde{\boldsymbol{s}}_0(t) = \boldsymbol{s}_{\theta'}(\boldsymbol{s}_t, \boldsymbol{x}_t, \boldsymbol{\mu}_{\boldsymbol{s}}, t), \tag{8}$$

which indicates that the subnetwork updating the single representation receives input from both the structure and the single representation as in the EvoFormer of AF2 (ref. 30). We note that the single representation was not used for updating the pair representation.

In theory, the ELBO-based denoising loss for the above DDPM model is defined per equation (9):

$$\mathcal{L}_{\theta,\theta'} = E_{t,\boldsymbol{x}_0}\left(\frac{\beta_t^2\bar{\alpha}_{t-1}}{2\sigma_t^2(1-\bar{\alpha}_t)^2}\left(||\boldsymbol{x}_\theta(\boldsymbol{x}_t,\boldsymbol{\mu}_x,t)-\boldsymbol{x}_0||^2+w_1||\boldsymbol{s}_{\theta'}(\boldsymbol{x}_t,\boldsymbol{s}_t,\boldsymbol{\mu}_s,t)-\boldsymbol{s}_0||^2\right)\right), \quad (9)$$

In our training of SCUBA-D, the overall structure-based loss $||\boldsymbol{x}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0||^2$ in equation (9) is replaced by several terms measuring structure recovery supplemented with adversarial losses (see below). The structure recovery terms comprised the FAPE loss ($FAPE_{DDPM}$), a distogram classification loss ($CE_{dist}$) and a covalent geometry violation loss ($loss_{violation}$). The $CE_{dist}$ is computed by predicting (using a learned subnetwork) the $C_\alpha$ distance map from the last layer pair representation of the denoising subnetwork, and comparing the predicted map with the actual map (to compute the loss as the sum of classification errors over all distances, the distance range within 20 Å is evenly divided into 64 bins and each bin is assigned a distance class). The $loss_{violation}$ is the mean squared deviations of bond lengths and bond angles from chemically allowed ranges.

The learning targets of the single representation $\boldsymbol{s}_0$ are based on ESM, a protein language model pretrained from massive samples of known natural protein sequences. The learning target of the 'compressed ESM' model was a compression of the ESM encoding of the native sequences into reduced dimensions; the learning target of the 'full ESM' model was the original ESM encoding the native sequences in full dimensions. The 'compressed ESM' representation was obtained by training an autoencoder of the full ESM representations of natural protein sequences. The autoencoder comprised three stacked transformer encoder blocks of dimensions 512, 256 and 512. The representation generated by the third block was used as a compressed representation of ESM.

## The adversarial losses

Two discriminator networks are used to provide adversarial losses to supplement the structure recovery losses in the denoising diffusion module. One is a graph-based packing discriminator, the other is a transformer-based backbone torsion correlation discriminator. These two discriminators evaluate the quality or physical plausibility of the denoised structure from two aspects: the first discriminator evaluates the local packing around individual residues, while the second discriminator evaluates the backbone conformations of contiguous peptide segments.

For the graph-based packing discriminator, the backbone structure is modeled as a graph, with nodes corresponding to amino acid residues and edges connecting spatially neighboring residues. The node features include the local backbone conformation represented by the 21 backbone torsional angles from $\phi_{i-3}$, $\psi_{i-3}$ and $\omega_{i-3}$ to $\phi_{i+3}$, $\psi_{i+3}$ and $\omega_{i+3}$, which are transformed with the sine and cosine functions. The edge features include distances between the N, $C_\alpha$, C and a virtual $C_\beta$ atoms of the two residues (these are similar to the edge features used by the sequence design program ProteinMPNN[27]). Besides these, the edge features also include a seven-component vector $\boldsymbol{O}_{ij}=[\boldsymbol{t}_{ij},\boldsymbol{q}_{ij}]$ describing relative orientations between the local coordinate frames constructed from three consecutive $C_\alpha$ atoms (that is, $C\alpha_{k-1}-C\alpha_k-C\alpha_{k+1}$, with $k\in\{i,j\}$). More specifically, $\boldsymbol{t}_{ij}$ is the normalized translation vector, and $\boldsymbol{q}_{ij}$ is the rotational quaternion from the local frame at $i$ to the local frame at $j$, both represented in the local frame at $i$.

The backbone torsion correlation discriminator is a transformer-based model, which encodes the backbone torsion angles ($\phi$,$\psi$) sequentially along the peptide chain by using global attention.

The discriminators were trained together with the DDPM with discriminator and generator losses. The discriminator losses are defined as shown in equation (10):

$$\mathcal{L}_{\mathcal{D}} = \frac{1}{2}\sum_{t=1}^{T}\lambda(t)E_{\boldsymbol{x}_0\sim q(x_0),\tilde{\boldsymbol{x}}_0(t)\sim p(\tilde{\boldsymbol{x}}_0(t),|,\boldsymbol{x}_0)}\left[(1-D(\boldsymbol{x}_0))^2+D(\tilde{\boldsymbol{x}}_0(t))^2\right], \quad (10)$$

where D (·) denotes the discriminator, and $\tilde{\boldsymbol{x}}_0(t)=\boldsymbol{x}_\theta(\boldsymbol{x}_t,t)$ is the generator. The generator losses ($loss_{gen}$) are defined as shown in equation (11):

$$\mathcal{L}_{\mathcal{G}} = \frac{1}{2}\sum_{t=1}^{T}\lambda(t)E_{\boldsymbol{x}_0\sim q(\boldsymbol{x}_0),\tilde{\boldsymbol{x}}_0(t)\sim p(\tilde{\boldsymbol{x}}_0(t),|,\boldsymbol{x}_0)}\left[[1-D(\tilde{\boldsymbol{x}}_0(t))]^2+||f(\tilde{\boldsymbol{x}}_0(t))-f(\boldsymbol{x}_0)||^2\right], \quad (11)$$

where f (·) denotes the values of the internal neuron nodes of the discriminator. Following the DDPM framework, these losses are weighted by $\lambda(t)=\sqrt{\bar{\alpha}_t}$.

## The training of SCUBA-D with structures of natural proteins

The available set of nonredundant natural protein domains (selected with the PISCES server[44] with a sequence identity cutoff of 70%) were split into training, validation and test sets in the following way: from each of the three fold classes—all-α, all-β and mixed αβ—domains of one randomly selected topology type were assigned to the validation set, and domains of another randomly selected topology type were assigned to the test set; the remaining domains were assigned to the training set. In this way, both the validation and the test sets contained domains of the three fold classes, while no pair of protein structures from two different sets shared the same topology type (according to the CATH 4.2 classification of protein structures[32]). The numbers of protein structures in the training, validation and test sets were 38,449, 1,130 and 572, respectively.

Our model has been trained by learning to denoise perturbed natural protein structures. To let the model learn how to recognize a diverse range of initial structures that the model will potentially run into in various downstream tasks, we used the following four different strategies to perturb natural structures to obtain initial structures.

S1: masking the structure information of a certain fraction of residues from a natural structure. The fraction of masked residues was randomly selected to be between 10% to 30% in the initial phase of training and increased to between 10% and 70% in the fine-tuning phase (see below) of training. The masked residues were multiple segments of 3 to 7 residues (sub-strategy S1–1), a single long segment (sub-strategy S1–2) or a set of residues that are spatial neighbors surrounding a randomly selected residue (sub-strategy S1–3). In the initial input structures, the masked residues take random positions (from a normal distribution with zero mean and 15 Å variance) and orientation (uniform distribution).

S2: adding random noise to the location and orientation of every residue in a natural structure. The noise vector added to the residue location was sampled from a normal distribution with zero mean and a variance of 3 Å. The orientation was sampled from the uniform distribution.

S3: adding random noises to the location and orientation of every secondary structure element in a natural structure. An entire secondary structure element is translated by a random vector sampled from a normal distribution with zero mean and a variance of 3 Å. The orientation is changed by interpolating between the orientation in the natural structure ($v_0$) and a random orientation ($v_1$) sampled from the uniform distribution, namely, as shown in equation (12):

$$v_{\text{perturbed}} = \frac{\sin(1-\gamma)\omega}{\sin\omega}v_0+\frac{\sin\gamma\omega}{\sin\omega}v_1, \quad (12)$$

where $\omega$ is the angle between the rotational axes, and the parameter $\gamma$ takes the value of 0.5 here. With the secondary structure elements perturbed, all residues in loops are treated as being masked.

S4: generating perturbed backbones by running stochastic dynamics (SD) simulations with SCUBA[9] starting from a natural structure. The

potential energy function used for the SD included the covalent and van der Waals energy terms but no other energy terms of SCUBA. The temperature of the simulations was set to 5 (in $k_BT$ unit) to encourage the sampling of diverse structures. For each training natural structure, the SD simulations were pre-performed for 100 ps and 20 structures evenly spaced in time were extracted from the last 70 ps of the simulation.

At each training step, a part of a maximum of 128 residues was cropped from a natural structure domain using the same approach as AF2 (the probabilities of using sequence-based cropping or using spatial location-based cropping is 0.5:0.5). The dataset for training the current SCUBA-D included approximately 38,000 cropped natural protein structures. An initial structure was generated by perturbing the cropped structure with one of the strategies among S1–S4 (with approximate probabilities 0.5:0.17:0.17:0.17). If S1 was chosen, the probabilities for choosing the sub-strategies S1–1 to S1–3 were approximately 0.33:0.33:0.33. The initial structures for strategies S1 to S3 were generated on the fly. If S4 was selected, one initial structure was randomly selected from the stored 20 structures produced by the pre-performed SD simulation.

The training was divided into two phases: an initial phase and a fine-tuning phase. The initial phase lasted for 50,000 training steps (one step corresponds to the processing of a batch of 144 training samples). The adversarial losses were not considered in this phase. The fine-tuning phase was started with the parameter values learned in the initial phase, which lasted 150,000 steps with the adversarial losses considered. The entire training took around 10 days on 24 V100 12 GB GPUs. PyTorch 1.11 software framework was used for construction and training the neural networks. More parameters about model training are summarized in Supplementary Table 5.

### Structure generation with SCUBA-D

During the inference stage, the number of diffusion steps used in inference (that is, backbone generation) was one-third of the number of steps used in training to increase the speed of structure generation. This was achieved by increasing the interval of $t$ by three folds in successive steps.

For generating sketched initial backbones according to a given architecture at the inference step, we used a protocol described in the SCUBA paper[9]. In this protocol, an architecture is defined by manually setting the types, approximate lengths and initial (approximate) positions of the secondary structure elements (the positions of a secondary structure elements were specified by the position of the N terminus and the direction of the N-to-C vector). Then a continuous peptide chain of correct covalent geometry was constructed by first generating the secondary structure segments at the specified spatial positions followed by generating loops with the kinematic loop closure algorithm to connect consecutive secondary structure segments. The loop lengths were automatically chosen to be the minimum lengths that can properly close the corresponding gaps. The computer program (named SCUBA-sketch) implementing the protocol is publicly available (Code availability), and several demos for using the program have been included in the SCUBA-D package.

### Designing and experimental characterizing specific proteins

Detailed descriptions of methods are provided in Supplementary Methods.

### Statistics and reproducibility

No statistical method was used to predetermine sample size. No data were excluded from the analyses unless explicitly specified.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Protein structures for training the models were downloaded from the PDB. The experimentally solved protein structures were deposited in the PDB under accession codes: 8K7Z (N1), 8K83 (N2), 8K84 (N3), 8KCJ (N7), 8KCK (N9), 8K8I (N14), 8KC4 (NA5), 8KA6 (NA7), 8KA7 (NB7), 8KC0 (NB8), 8KAC (NX1), 8KC1 (NX5), 8K7M (T01), 8KDQ (T03), 8WX8 (T09), 8KC8 (T11) and 8WWC (120–4). We referenced the structures 2ZDO and 4G0N from the PDB for the design of heme-binding proteins and Ras-binding proteins, respectively. The amino acid sequences and encoding DNA sequences of the experimentally examined proteins are available in Supplementary Tables 6–10 and Supplementary Data 1–3. The complete lists of proteins for training and testing the models, the data of experimental results (SEC, multi-angle light scattering, $^{15}N$-$^{1}H$ HSQC NMR, ITC, CD, validation reports for experimentally solved protein structures) and all in silico experimental results are available from Zenodo via https://doi.org/10.5281/zenodo.10911626 (ref. 45). Source data are provided with this paper.

## Code availability

Executable computer programs and source codes of SCUBA-D (version 1.0) and SCUBA-sketch (version 1.0) are publicly available from Zenodo via https://doi.org/10.5281/zenodo.10947360 (ref. 46) and can be freely used for noncommercial purposes. The source codes for SCUBA-D are also available from GitHub at https://github.com/liuyf020419/SCUBA-D.git/.

## References


43. Jing, B., Eismann, S., Suriana, P., Townshend, R. J., Dror, R. Learning from protein structure with geometric vector perceptrons. In *Proc. International Conference on Learning Representations* (ICLR, 2021).
44. Wang, G. & Dunbrack, R. L. Jr PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589–1591 (2003).
45. Wang, S. Source data for manuscript: de novo protein design with a denoising diffusion network independent of pre-trained structure prediction models. *Zenodo* https://doi.org/10.5281/zenodo.10911626 (2024).
46. Wang, S. De novo protein design with a denoising diffusion network independent of pre-trained structure prediction models. *Zenodo* https://doi.org/10.5281/zenodo.10947360 (2024).


## Acknowledgements


We thank the staff from the BL18U1 and BL19U1 beamlines of the National Facility for Protein Science in Shanghai for their assistance during crystallographic data collection. We also thank X. Hu, R. Wu and L. Zhang for their help with experimental techniques, as well as M. Lv and H. Yu for their help with crystal collection. This work was supported by the National Key R&D Program of China (2022YFA1303700 to H.L. and 2022YFF1203100 to Q.C.), National Natural Science Foundation of China (T2221005, 92253302 and 22177107 to H.L.; 32371487 and 32171411 to Q.C.), CAS Strategic Priority Research Program (XDB0500201 to H.L.), CAS Project for Young Scientists in Basic Research (YSBR-072 to Q.C.), Anhui Provincial Natural Science Foundation (2308085J01 to Q.C.) and Research Funds of Center for Advanced Interdisciplinary Science and Biomedicine of IHM (QYPY20230035 to Q.C.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.


## Author contributions

Y.L. developed computational models and codes with the assistance of L.C. S.W. carried out the experimental work with the help of J.D., X.W. and Y.W. L.W., F.L. and C.W. helped with analysis of crystal structural data. J.Z. collected the NMR data. S.W. participated in the discussion. H.L. and Q.C. supervised the project. H.L., Y.L., S.W. and Q.C. wrote the paper.

## Competing interests
The authors declare no competing interests.

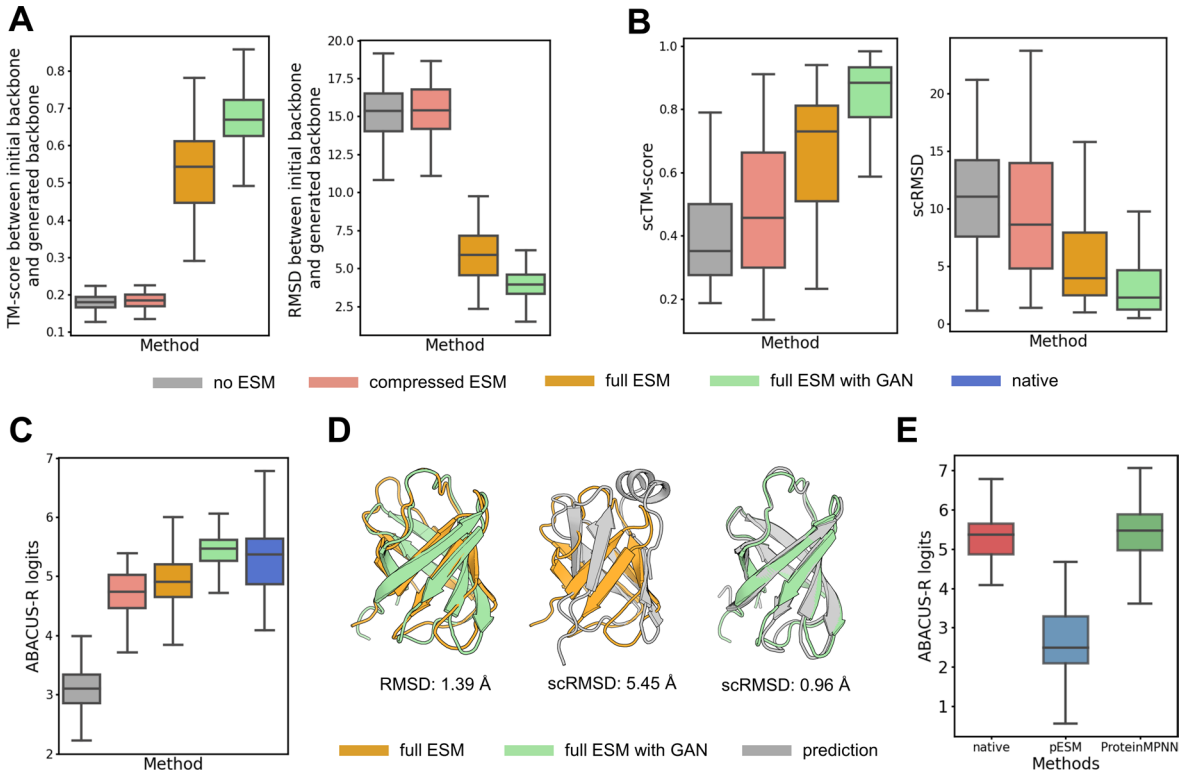
## Additional information
**Extended data** is available for this paper at https://doi.org/10.1038/s41592-024-02437-w.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02437-w.

**Correspondence and requests for materials** should be addressed to Quan Chen or Haiyan Liu.

**Peer review information** *Nature Methods* thanks Arne Elofsson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Arunima Singh, in collaboration with the *Nature Methods* team.

https://doi.org/10.1038/s41592-024-02437-w

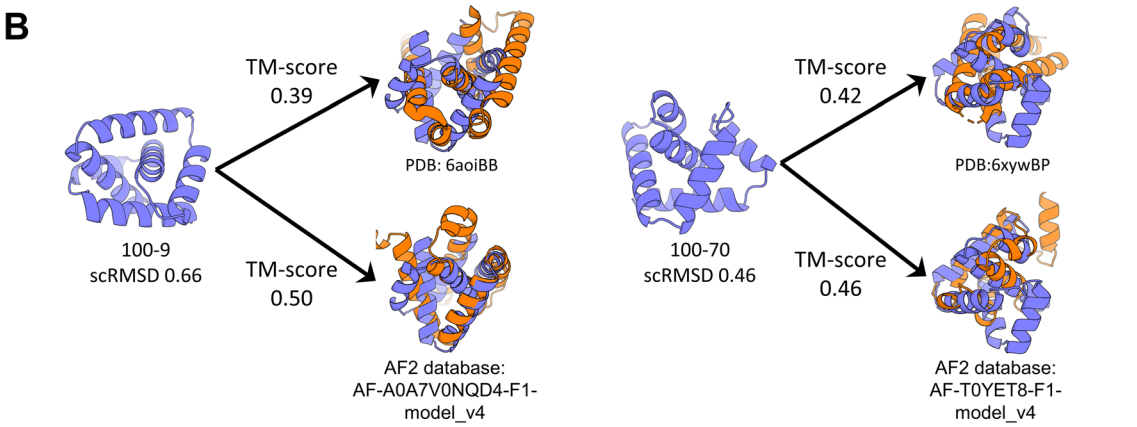


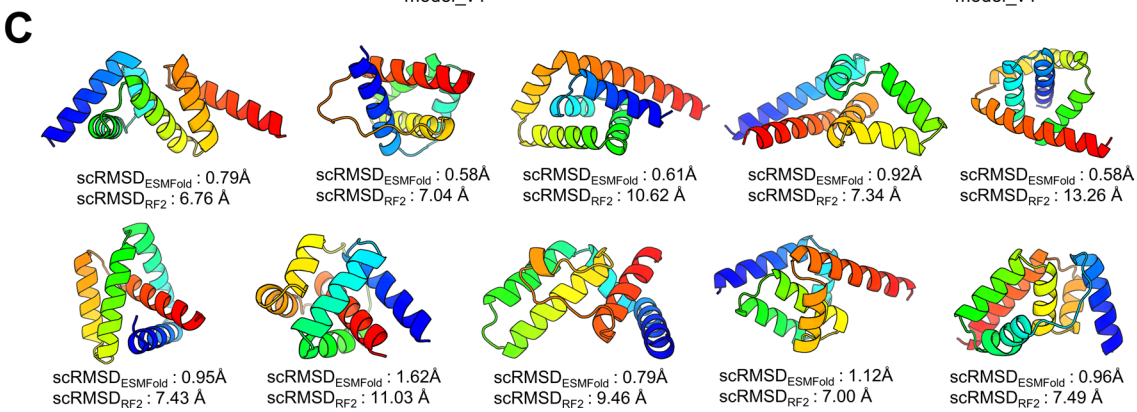

**Extended Data Fig. 1 | Evaluation of the variant models 'no ESM', 'compressed ESM', 'full ESM' and 'full ESM with GAN'.** (**a**) Distributions of TM-scores and RMSDs between the initial natural backbones and the denoised backbones generated by the variant modes. For each model, 75 protein backbones were generated by considering 3 independent 'denoising' runs from each of 25 initial natural backbones. (**b**) Distributions of scTM-scores and scRMSDs between the denoised backbones and the AlphaFold2-predicted structures for amino acid sequences designed (with ABACUS-R) on corresponding denoised backbones. (**c**) Distributions of per-residue ABACUS-R logits scores of amino acid sequences designed by ABACUS-R for the various denoised backbones and for the initial natural backbones. Larger logits scores indicate better compatibility between the designed sequences and the corresponding backbone structures. (**d**) Left: backbones 'denoised' with the 'full ESM' model (orange) and the 'full ESM with GAN' model (green) from the same initial natural backbone 1e1qA01 (CATH domain ID); the RMSD between the two denoised backbones is indicated. Middle: the backbone denoised with the 'full ESM' model (orange) superimposed with the AlphaFold2-predicted structure (gray) for the amino acid sequence designed on this backbone by ABACUS-R; the corresponding scRMSD is indicated. Left: the backbone denoised with the 'full ESM with GAN' model (green) superimposed with the AlphaFold2-predicted structure (gray) for the amino acid sequence designed on this backbone by ABACUS-R; the corresponding scRMSD is indicated. (**e**) The distributions of the ABACUS-R logits of different amino acid sequences for 25 natural backbones. The pESM sequences were obtained by projecting the single representation parts from the SCUBA-D output using a residue type classifier network of ESM. The boxplots in A to C and E show median, interquartile range, and minimum and maximum values excluding outliers (>1.5 times the interquartile range beyond the box) with the sample sizes being 75 (for the denoised backbones) or 25 (for the natural backbones).

https://doi.org/10.1038/s41592-024-02437-w

**A**

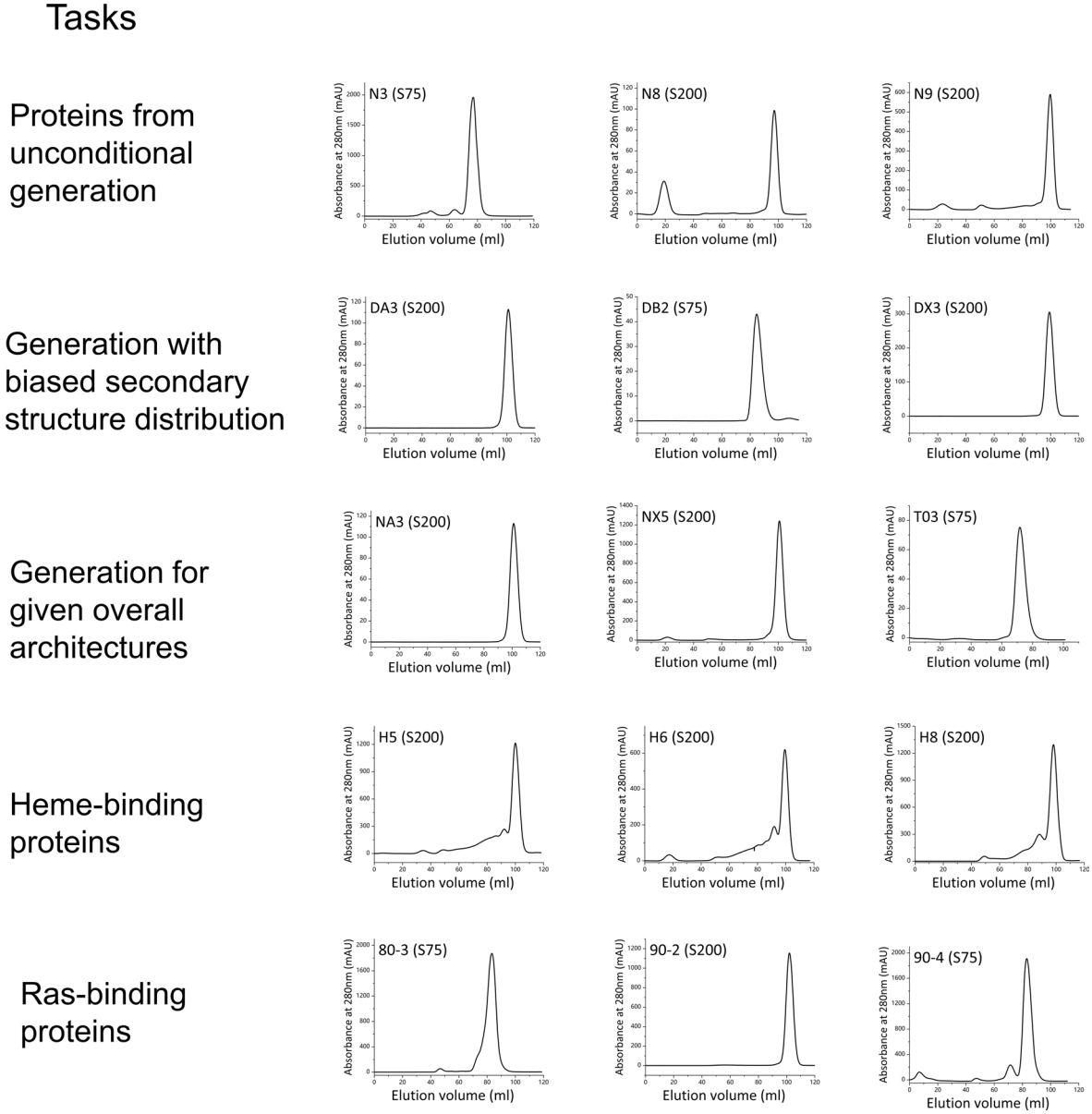
| | method | chain length | mean scRMSD (# < 2.5 Å) | mean Highest TM-score to PDB (# <0.5) | mean number of clashed residue per backbone |
|---|---|---|---|---|---|
| fine-tuned model | RFdiffusion | 100 | 0.72 (97) | 0.69 (2) | 1.53 |
| | | 200-400 | 3.28 (204) | 0.52 (125) | 13.25 |
| freshly-trained model | SCUBA-D | 100 | 0.78 (99) | 0.62 (18) | 2.37 |
| | | 200-400 | 3.46 (194) | 0.51 (154) | 7.32 |
| | Chroma | 100 | 3.03 (60) | 0.59 (17) | 0.41 |
| | | 200-400 | 8.15 (66) | 0.52 (125) | 0.41 |
| | FrameDiff | 100 | 2.44 (69) | 0.70 (6) | 25.88 |
| | | 200-400 | 9.32 (45) | 0.57 (106) | 73.78 |
| | Genie | 100 | 2.36 (73) | 0.56 (25) | - |
| | | 200-400 | - | - | - |
| | ProteinSGM | 100 | 3.72 (50) | 0.65 (14) | - |
| | | 200-400 | - | - | - |



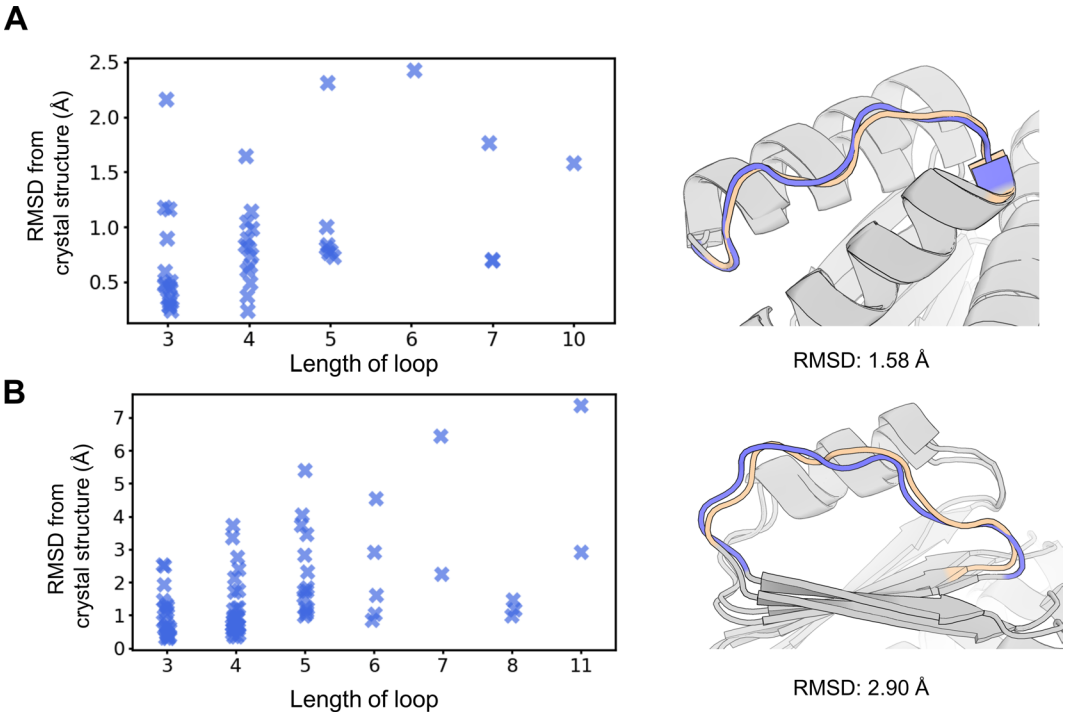

**Extended Data Fig. 2 | Comparasions between SCUBA-D and other DDPM models for unconditional backbone generation.** (**a**) Averaged metrics of various models. For each method, the averages over two groups (one group comprised 100 backbones of 100 residues in chain length and the other group comprised 300 backbones of 200 to 400 residues in chain length) are reported, with data in the parentheses reporting the total number of backbones with scRMSD below 2.5 Å or the total number of backbones of high overall structural novelty (the highest TM-score to PDB below 0.5). (**b**) Two example backbones of 100 residues (100-9 and 100-7) generated by SCUBA-D without condition and with their highest TM-scores to both PDB and AlphaFold2 database below or equal to 0.5. The generated backbones (in blue) and their superimpositions with the corresponding structures from PDB or AlphaFold2 database (in salmon) are shown. The respective TM-scores and PDB IDs (with chain IDs) are indicated. Here the scRMSD of a generated backbone was determined as the RMSD of the backbone from the AF2 predicted structure for the amino acid sequence designed (here with ProteinMPNN) for that backbone. (**c**) The structures of ten example backbones with RosettaFold2-based scRMSDs above 6.0 Å. Both the ESM prediction-based scRMSDs and the RosettaFold2 prediction-based scRMSDs are indicated.

https://doi.org/10.1038/s41592-024-02437-w



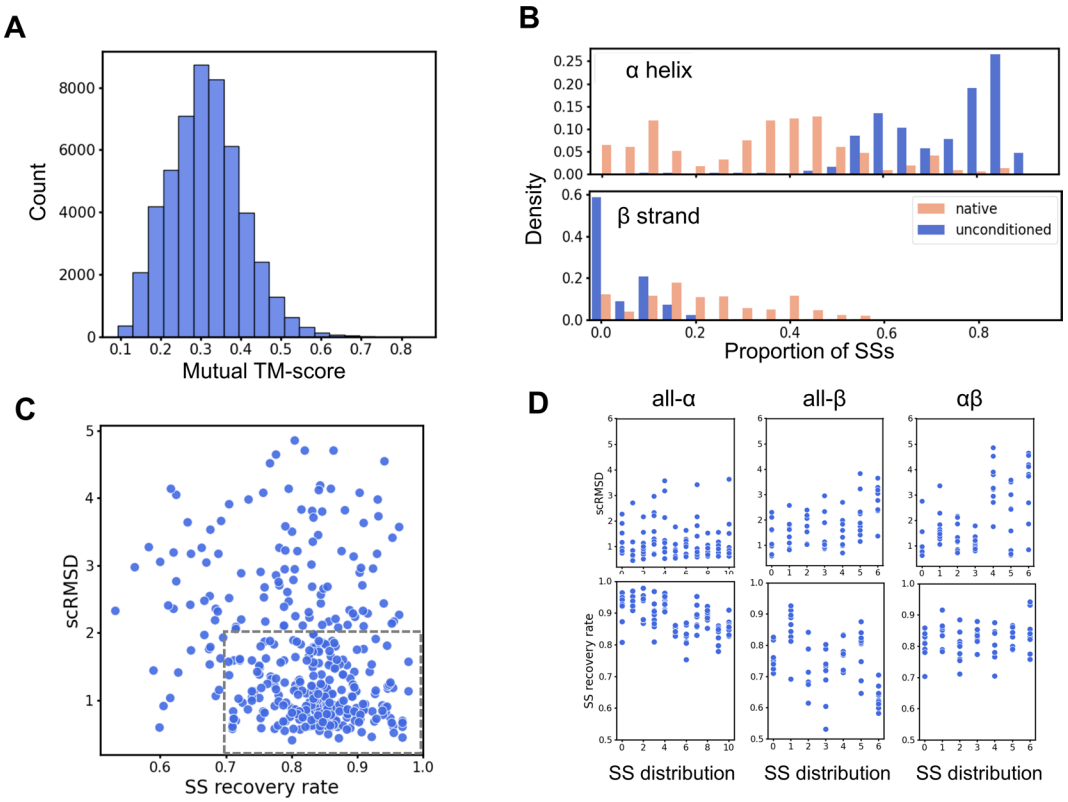

**Extended Data Fig. 3 | Example results of size-exclusion chromatography (SEC) experiments.** Proteins designed in the five different tasks as indicated were analyzed. For each task, three example results are shown in the same row. The protein IDs and the types of the SEC columns are indicated.

https://doi.org/10.1038/s41592-024-02437-w



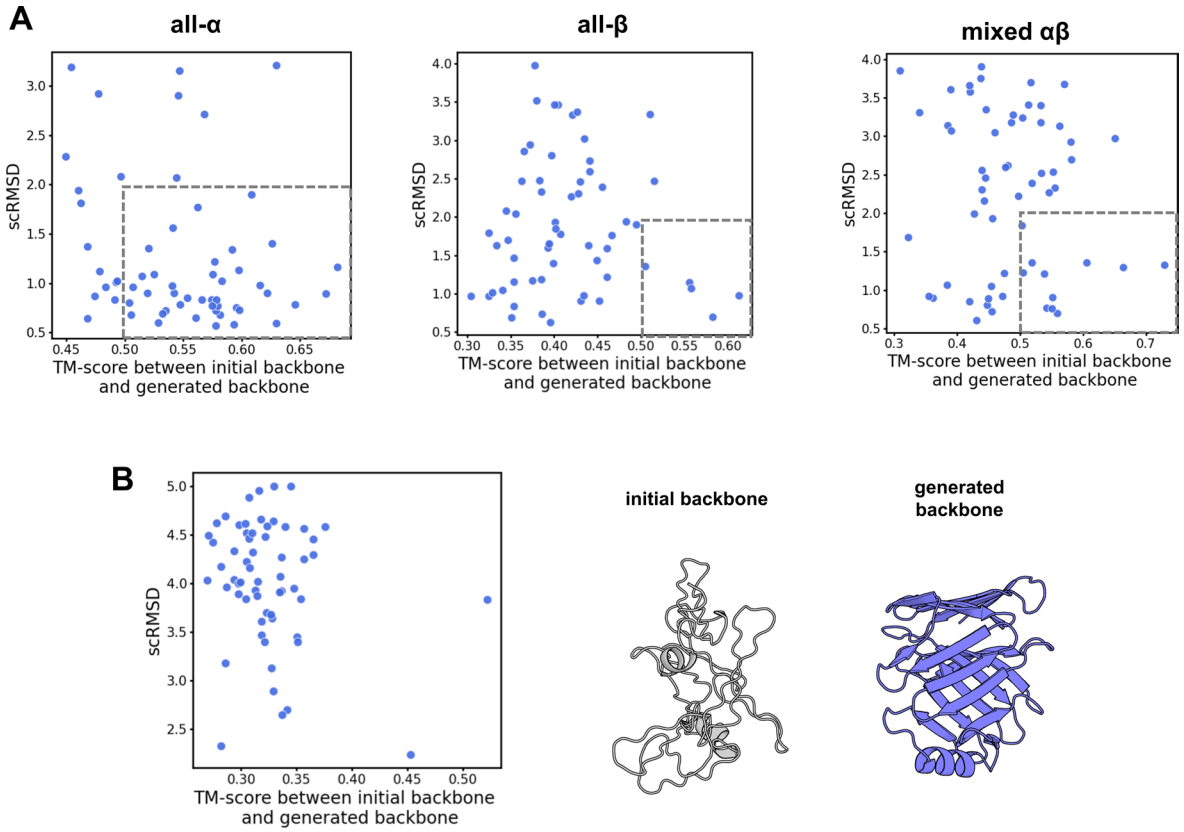

**Extended Data Fig. 4 | The deviations between the loops in designed structures and in solved crystal structures.** (**a**) The RMSDs between the loops. The analysis included the 6 crystal structures obtained for proteins of backbones generated by SCUBA-D without condition. Each point corresponds to a loop, with the loops grouped according to their lengths and those of the same length displayed in the same column. The RMSDs were calculated by superimposing the flanking secondary structure segments for a pair of compared loops. An example showing the superimposed structures with the indicated RMSD between designed loop (blue) and corresponding crystal structure (orange). (**b**) The same as A, but for the 6 experimentally determined structures of proteins generated for particular architectures.

https://doi.org/10.1038/s41592-024-02437-w



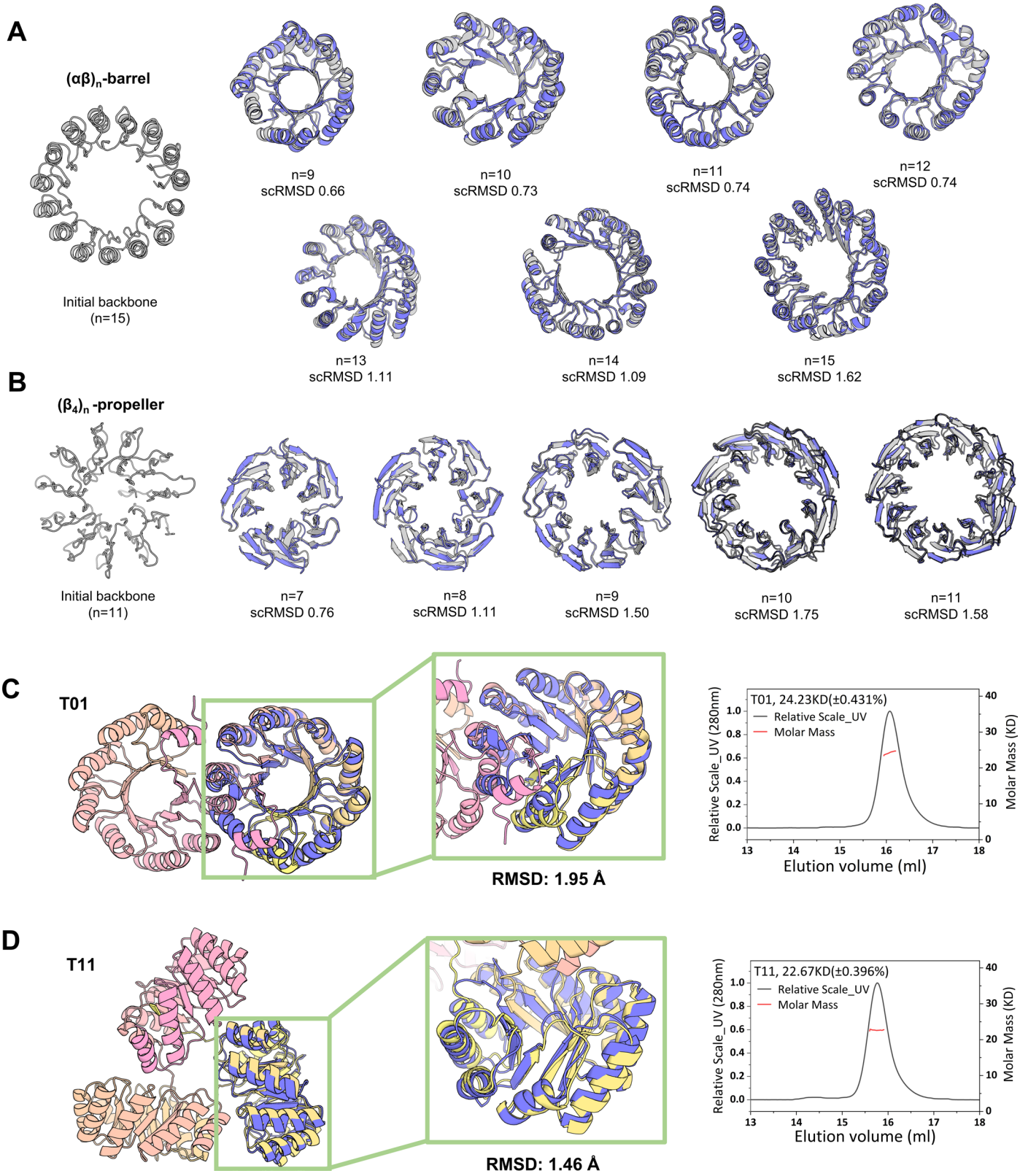

**Extended Data Fig. 5 | Protein backbone generation without condition or with biased secondary structure (SS) distributions.** (**a**) The distribution of the mutual TM-scores between the set of backbones unconditionally generated by SCUBA-D. (**b**) Histograms of the proportions of residues in the α helix state (upper panel) and of residues in the β strand state (lower panel) for the set of unconditionally generated backbones with SCUBA-D (blue) and for a set of natural protein structures (salmon), which comprised PDB structures of resolutions higher than 2.0 Å, of mutual sequence identities below 40%, and of 100 to 500 residues in length. The proportions were calculated on individual backbones. The histograms represent the normalized frequencies of backbones with proportions in specific bins. (**c**) Scattering plot of the recovery rates of the input secondary structure (SS) states versus the scRMSDs for the set of 225 backbones generated using the 25 input structures. Each input structure was composed according to the SS distribution of a natural backbone. The gray box indicates the region with scRMSD < 2.0 Å and SS recovery rate > 70%. (**d**) The scRMSDs of the backbones generated with biased SS distributions and the SS recovery rate. For each SS distribution, 9 backbones were generated and evaluated, one data point in the plots corresponding to one designed backbone. The results for three different classes of SS distributions (all-α, all-β, and mixed αβ) are displayed in different plots. Within each plot, results biased towards the same SS distribution are numbered the same and displayed in the same column. Results for different SS distributions were arranged from left to right in an ascending order of the corresponding chain lengths.

https://doi.org/10.1038/s41592-024-02437-w




**Extended Data Fig. 6 | Backbone generation with skected input structures.** (**a**) Example scattering plots of scRMSD versus TM-score to initial structure for the backbones generated from initial structures 'sketched' according to three architectures of different natural proteins. The examples were of different fold classes (all-α, all-β, and mixed αβ). For each architecture, backbones were generated by applying SCUBA-D to 60 independently 'sketched' input structures. The dashed boxes indicate regions with scRMSDs < 2.0 Å and the TM-scores to initial backbones > 0.5. (**b**) An example for which no generated backbone for the particular architecture meet the criteria of scRMSD < 2.0 Å and TM-score > 0.5. Left: the scattering plot of scRMSD versus the TM-score to initial and backbone for the architecture. Middle: an example of the initial backbone. Right: an example generated backbone.

https://doi.org/10.1038/s41592-024-02437-w

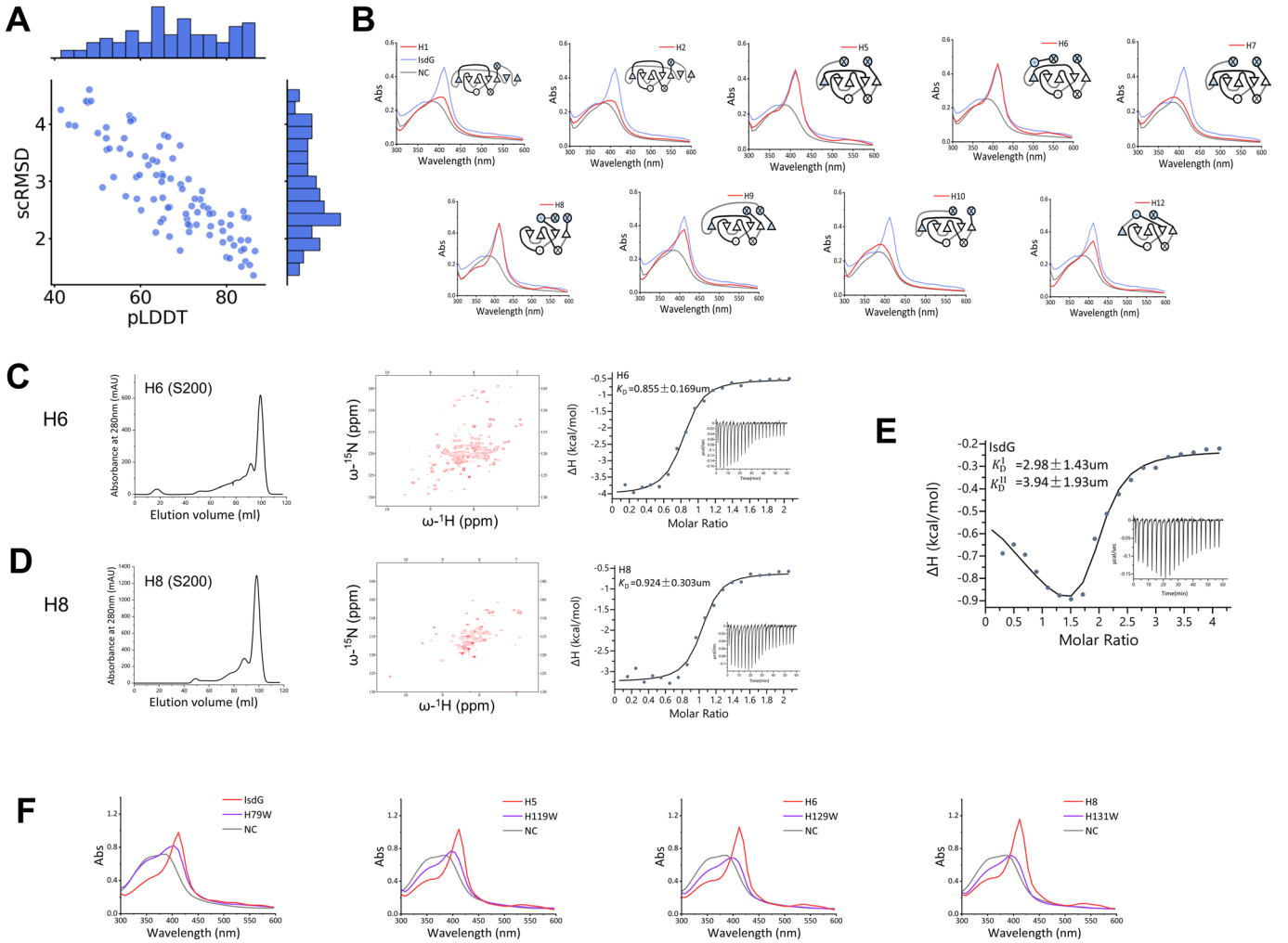



**Extended Data Fig. 7 | See next page for caption.**

https://doi.org/10.1038/s41592-024-02437-w

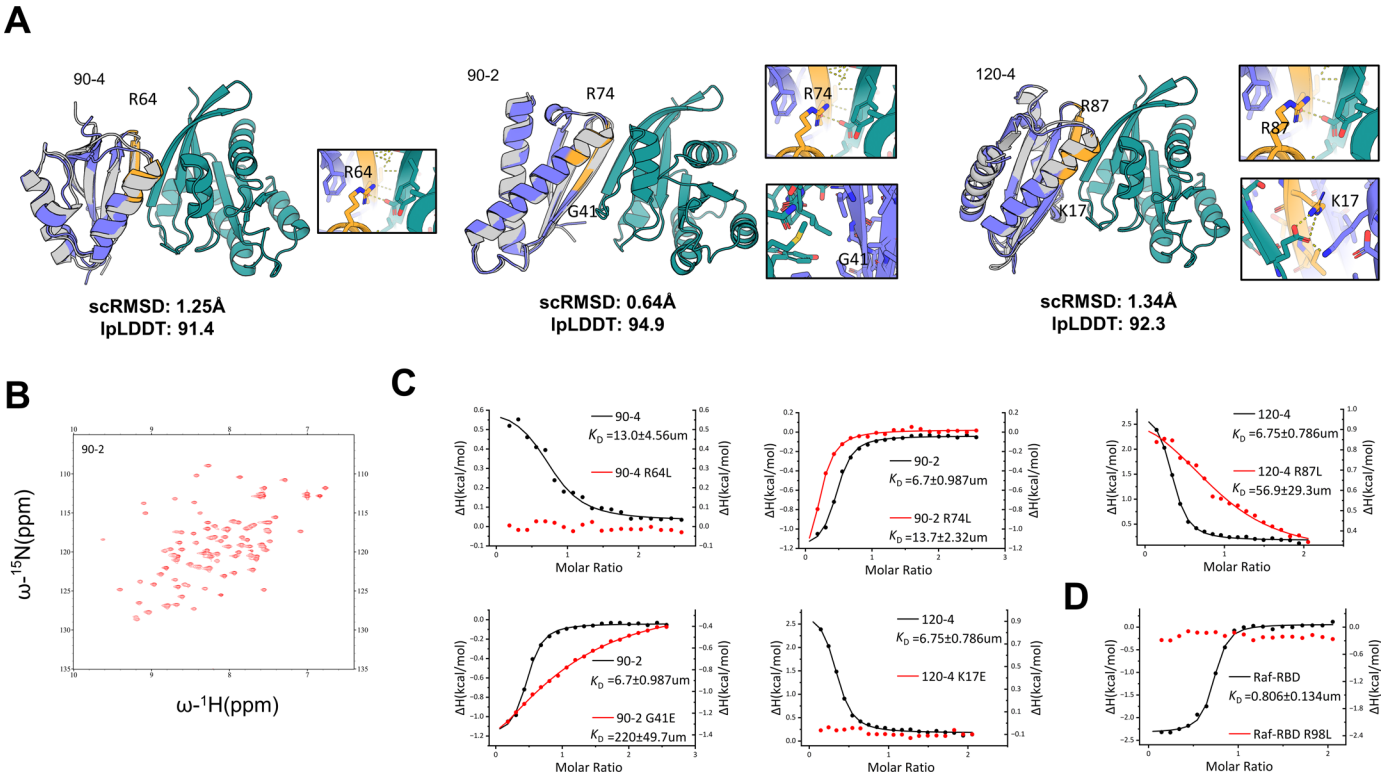
**Extended Data Fig. 7 | Designing proteins of the (αβ)n-barrel and the (β4) n-propeller architectures.** (**a**) Left: an example initial structure 'sketched' according to the $(\alpha\beta)_{15}$-barrel architecture. Right: example backbones (blue) generated for the $(\alpha\beta)_n$-barrel architectures superimposed with structures predicted by AlphaFold2 (gray) for amino acid sequences designed for these backbones with ProteinMPNN. The scRMSDs of the superimpositions are indicated. For each value of the repeat number n from 9 to 15, one example is shown. (**b**) The same as A, but for the $(\beta_4)_n$-propeller architectures with n ranging from 7 to 11. (**c**) Left: the crystal structure (gold and salmon) and the designed backbone (blue) of the designed $(\alpha\beta)_9$-barrel protein T01. The crystal structure presents a domain-swapped dimer, with the monomers colored differently. The designed backbone is superimposed with one of the monomers. Right: the results of SEC (black curve) and static light scattering (red curve) experiments on T01, which indicate that the protein exists in the monomeric state in solution. (**d**) Left: the crystal structure (gold, yellow, and salmon) and the designed backbone (blue) of the designed $(\alpha\beta)_9$-barrel protein T11. The crystal structure presents a domain-swapped trimer, with the monomers colored differently. The designed backbone is superimposed with one of the monomers. Right: the results of SEC (black curve) and static light scattering (red curve) experiments on T11, which indicate that the protein exists in the monomeric state in solution.

https://doi.org/10.1038/s41592-024-02437-w



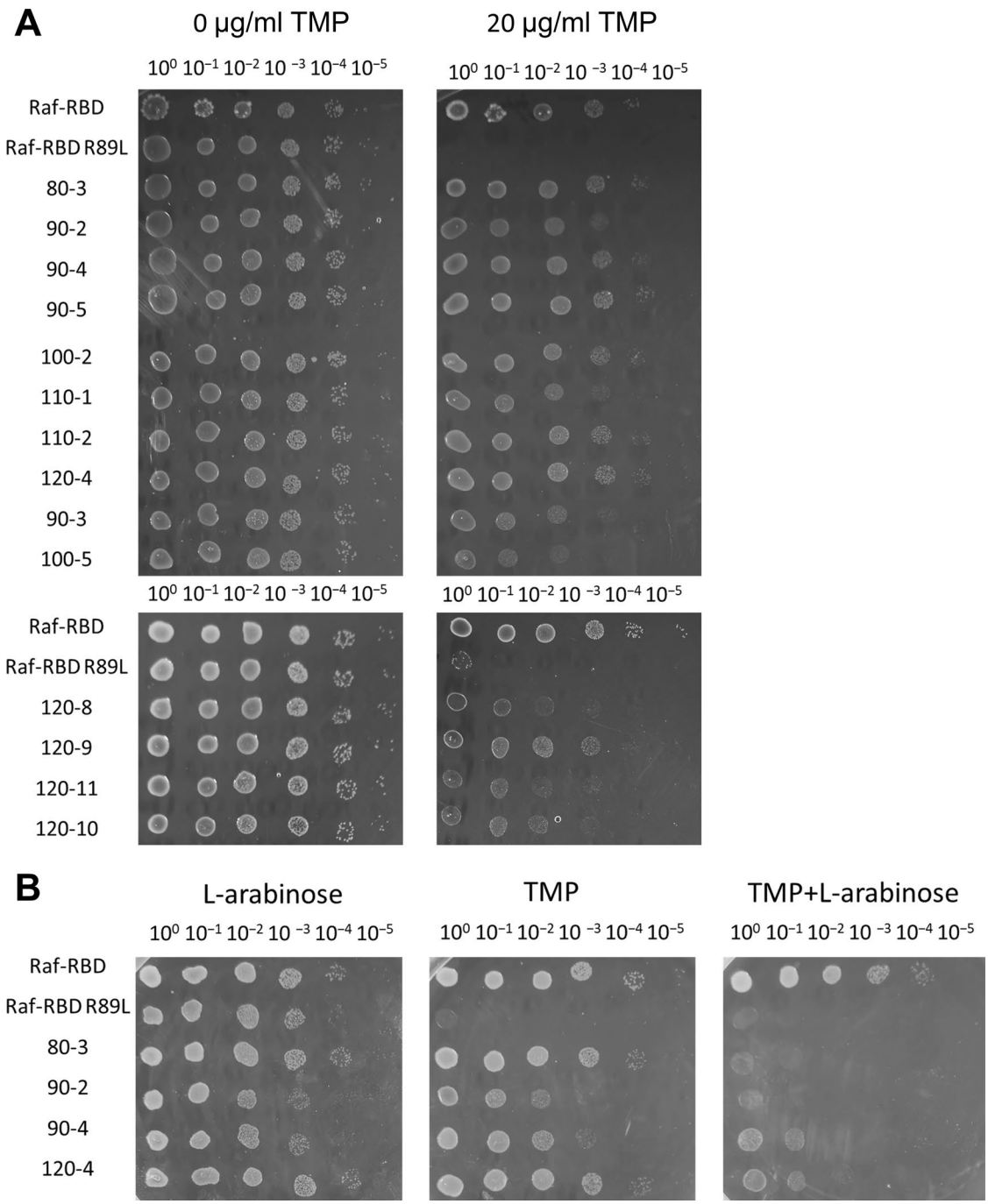

**Extended Data Fig. 8 | Designed heme-binding proteins.** (**a**) Scattering plot and histograms of the scRMSD and pLDDT scores of the designed heme-binding backbones. Structure predictions with AlphaFold2 were performed for amino acid sequences designed with the ABACUS-R program. (**b**) UV-Visible absorbance spectra of 9 designed heme-binding proteins are shown with the topology diagrams of the corresponding proteins. 'NC' represents negative control. 'IsdG' represents the natural iron-regulated surface determinant G protein which served as a positive control. Heme binding is indicated by the presence of the peak around 412 nm. (**c**) Experimental characterizations of the designed heme-binding protein H6. Left: SEC result. Middle: NMR $^{15}$N-$^{1}$H HSQC spectrum. Right: ITC measurements on heme binding. (**d**) The same as **C**, but for H8. (**e**) The result of ITC experiments measuring the $K_D$ values of heme binding by the natural protein IsdG. (**f**) UV-Visible absorbance spectra showing the impacts of mutating the iron-coordinating histidine residues in the natural protein and the designed heme-binding proteins. Each panel shows the spectrum of a mutated protein together with the spectra of the corresponding original protein and of a non-heme-binding negative control protein (labeled as 'NC').

https://doi.org/10.1038/s41592-024-02437-w




**Extended Data Fig. 9 | The designed structures and experimental characterizations of the Ras-binding proteins 90-4, 90-2 and 120-4.** (**a**) The designed proteins (90-4, 90-2 and 120-4, colored in blue) are correspondingly superimposed with the predicted structures (gray) in complex with Ras (green). For each designed protein, the residues to be mutated is shown with its surrounding residues in the predicted structure next to the overall superimposition. The scRMSD and ligand pLDDT are indicated. (**b**) NMR $^{15}$N-$^{1}$H HSQC spectrum of Ras-binding proteins 90-2. (**c**) The results of ITC measurements on the Ras binding of 90-4, 90-2 and 120-4 and their mutated variants. (**d**) The results of ITC measurements on the Ras binding of Raf-RBD and the mutated variant of Raf-RBD (R89L).

https://doi.org/10.1038/s41592-024-02437-w




**Extended Data Fig. 10 | Assessing the designed Ras-binding proteins with the dihydrofolate reductase (DHFR)-based protein complementarity analysis assay.** (**a**) The protein complementarity analysis results on 14 designed Ras-binding proteins. In these experiments, the peptide chain of DHFR is split into two parts. Ras and the protein to be assessed were separately fused with each part. Bacterium cells expressing the two fused peptides were diluted to different levels of concentrations and tittered on media containing different levels of trimethoprim (TMP), which can inhibit the endogenous DHFR activity of the cells. Possible binding between Ras and the protein to be assessed was detected through the resistance of the bacterium cells to the growth inhibition by TMP. The label 'Raf-RBD' represents the Ras-binding domain of Raf, which served as a positive control. The label 'Raf-RBD R89L' represents a mutant with abolished Ras binding activity, which served as a negative control. The stronger TMP resistance (relative to the negative control) exhibited by the cells expressing fusion peptides of the designed proteins indicated that the designed proteins examined here can bind Ras. (**b**) Results of competitive DHFR-PCA analysis of 4 designed proteins. In the experiments examining a designed protein, cells co-expressing isolated Raf-RBD and the DHFR-PCA system for the designed protein were analyzed. If the designed protein and Raf-RBD share binding sites on Ras, the expression of Raf-RBD, which is induced by L-arabinose, will lead to the competitive inhibition of the Ras binding of the designed protein, detected as reduced resistance to TMP.

nature portfolio

Corresponding author(s): QUAN CHEN, HAIYAN LIU

Last updated by author(s): Jun 18, 2024

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

**Data collection**

PISCES server (http://dunbrack.fccc.edu/Guoli/pisces_download.php) was used to select non-redundant data with a resolution cutoff of 4.0 Angstrom and a sequence-identity cutoff of 70%. CATH server (http://cathdb.info/wiki?id=data:index, version 4.3) was used to assign the class of architecture for the downloaded protein structures.
The software framework for constructing and training the neural network models was PyTorch 1.11. Executable computer programs and source codes of SCUBA-D (version 1.0) and SCUBA-sketch (version 1.0) are publicly available from Zenodo at https://doi.org/10.5281/zenodo.10947360 and can be freely used for non-commercial purposes. The source codes for SCUBA-D are also available from GitHub at https://github.com/liuyf020419/SCUBA-D.git.

**Data analysis**

We used Python3.8, Numpy, Pandas, Matplotlib and Seaborn to analyze data. The protein structures are visualized with PyMOL version 1.8. For structure prediction, source code for the AlphaFold model, trained weights, and inference script were used (https://github.com/deepmind/alphafold). We applied TMscore (https://zhanggroup.org/TM-score) to superimpose the overall protein structure backbone. ProteinMPNN software (https://github.com/dauparas/ProteinMPNN) and ABACUS-R software (https://doi.org/10.24433/CO.3351944.v1) were used to design the amino acid sequences for the given backbones. We used Foldseek software (https://search.foldseek.com) to search against PDB and AlphaFold2 prediction database.

NMR data were processed using the software NMRDraw/NMRPipe (Version 8.2) and SPARKY 3.115. Crystallographic data for structures N1, N2, N3, N7, N9, N14, NA5, NA7, NB7, NB8, NX1, NX5, T01, T03, T09, T11, and 120-4 were processed using XDS (Version Feb 5, 2021). The designed structures served as search models for molecular replacement with PHENIX (v1.20.1-4487), and the final structures were refined using PHENIX (v1.20.1-4487). Structure figures were made with PyMOL version 1.8.

Isothermal titration calorimetry data were processed using Origin software provided by Microcal(Version 3.8.0). Circular dichroism spectra

data were processed, and secondary structure contents were estimated using the built-in software suite ProData Viewer v4.5 and Deconvolution v2.1, respectively. Multi-angle light scattering data were processed using the ASTRA software (v7.0.1) provided by Wyatt Technology, with graphical representation achieved using Origin 2020. The UV-visible spectroscopy data analysis was performed using the SoftMax® Pro software (v7.1.1) provided by Molecular Devices, and graphical representation was achieved using Origin 2020.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Protein structures for training the models have been downloaded from the Protein Data Bank. The coordinates and structure files of the designed proteins have been deposited in the Protein Data Bank under the following accession codes: 8k7z (N1), 8k83 (N2), 8k84 (N3), 8kcj (N7), 8kck (N9), 8k8i (N14), 8kc4 (NA5), 8ka6 (NA7), 8ka7 (NB7), 8kc0 (NB8), 8kac (NX1), 8kc1 (NX5), 8k7m (T01), 8kdq (T03), 8wx8 (T09), 8kc8 (T11), and 8wwc (120-4).
We referenced the structures 2zdo and 4g0n from the Protein Data Bank for the design of heme-binding proteins and Ras-binding proteins, respectively.
The complete lists of proteins for training and testing the models, the experimentally source data (size-exclusion chromatography data, multi-angle light scattering data, 15N-1H HSQC NMR data, ITC source data, circular dichroism data, validation reports for experimentally solved protein structures) and all in silico experimental results are available from Zenodo at https://doi.org/10.5281/zenodo.10911626.

## Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

| | |
|---|---|
| Reporting on sex and gender | N/A |
| Population characteristics | N/A |
| Recruitment | N/A |
| Ethics oversight | N/A |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | Sample sizes were chosen prior to the experiments. Sample sizes were chosen to be sufficiently large for estimating distributional properties, such as predicting pLDDT and scRMSD distribution, the distribution of the secondary structure. Then, the sample sizes were arbitrarily determined by experimenting (rather than through statistical tests). We note that the determined sample sizes are large enough to draw meaningful conclusions from the experiments while leading to affordable computational and experimental costs.<br>From the unconditionally generated backbones, 16 proteins were arbitrary picked for biological experiments, out of which 12 led to successful protein expression in E. coli. We obtained protein crystals for 7 proteins and solved 6 X-ray structures (with PDB IDs: 8k7z (N1), 8k83 (N2), 8k84 (N3), 8kcj (N7), 8kck (N9), and 8k8i (N14)).<br>From backbones generated with biased secondary structure distributions, 17 proteins were arbitrary picked, with 12 resulting in successful protein expression in E. coli. We obtained 3 protein crystals but were unable to acquire X-ray diffraction data suitable for structure determination.<br>Among the 37 proteins designed for the backbones generated from sketched inputs, 29 led to successful protein expression in E. coli. We obtained and solved X-ray structures for 10 proteins (with PDB IDs: 8kc4 (NA5), 8ka6 (NA7), 8ka7 (NB7), 8kc0 (NB8), 8kac (NX1), 8kc1 (NX5), 8k7m (T01), 8kdq (T03), 8wx8 (T09), and 8kc8 (T11)). Additionally, two design proteins were measured by Multi-angle light scattering experiments.<br>From designed heme-binding proteins filtered by in silico criteria, 12 designs were arbitrarily picked for biological experiments, resulting in 9 successful protein expressions in E. coli. UV-visible spectra experiments indicated 5 of them could bind heme. The affinity for heme binding of |

3 designed heme-binding proteins was verified by ITC experiments. The folding of these proteins were confirmed by NMR 15N-1H HSQC spectrum.
From designed Ras-binding proteins filtered by in silico criteria, 30 designs were arbitrarily picked for biological experiments, among which 14 showed detectable binding in the dihydrofolate reductase-based protein complementarity analysis (PCA) assay. Competitive PCA analysis on 4 designed proteins revealed that co-expressed Raf competitively inhibited the binding of the designed proteins with Ras. The affinity of three designed proteins and their mutants with RAS was verified by ITC experiments. Additionally, we obtained a complex crystal of the designed protein binding to RAS and solved its X-ray structure (PDB ID: 8wwc (120-4)).

Data exclusions: No data were excluded.

Replication: Protein expression and solubility was tested once or twice. All attempts to replicate expression and solubility screening experiments for further experimental characterization were successful. The X-ray structures were determined based on single crystals using standard procedures which have internal statistical validations.The ITC and UV-visible spectra experiments were repeated two or three times. Protein-fragment complementation assay (PCA) assay was tested twice. We performed multi-angle light scattering experiments once, as the signal-to-noise ratio of light scattering detector (connected with well flushed SEC columns) in a multi-angle light scattering machine was sufficiently high for determining the molar masses with acceptably low uncertainty in one experiments.

Randomization: There was no randomized sample allocation in this work. All tested protein designs received identical treatment.

Blinding: Blinding is not relevant to our study because there is no group allocation.

# Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description: *Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*

Research sample: *State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*

Sampling strategy: *Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*

Data collection: *Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*

Timing: *Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*

Data exclusions: *If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*

Non-participation: *State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*

Randomization: *If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

# Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description: *Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*

Research sample: *Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*

Sampling strategy: *Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*

| | |
|---|---|
| Data collection | *Describe the data collection procedure, including who recorded the data and how.* |
| Timing and spatial scale | *Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken* |
| Data exclusions | *If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.* |
| Reproducibility | *Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.* |
| Randomization | *Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.* |
| Blinding | *Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.* |

Did the study involve field work? ☐ Yes ☐ No

## Field work, collection and transport

| | |
|---|---|
| Field conditions | *Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).* |
| Location | *State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).* |
| Access & import/export | *Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).* |
| Disturbance | *Describe any disturbance caused by the study and how it was minimized.* |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

### Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Antibodies

| | |
|---|---|
| Antibodies used | *Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.* |
| Validation | *Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.* |

## Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

| | |
|---|---|
| Cell line source(s) | *State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.* |

| | |
|---|---|
| Authentication | *Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.* |
| Mycoplasma contamination | *Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.* |
| Commonly misidentified lines (See ICLAC register) | *Name any commonly misidentified cell lines used in the study and provide a rationale for their use.* |

## Palaeontology and Archaeology

| | |
|---|---|
| Specimen provenance | *Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.* |
| Specimen deposition | *Indicate where the specimens have been deposited to permit free access by other researchers.* |
| Dating methods | *If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.* |

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

| | |
|---|---|
| Ethics oversight | *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.* |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

| | |
|---|---|
| Laboratory animals | *For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.* |
| Wild animals | *Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.* |
| Reporting on sex | *Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.* |
| Field-collected samples | *For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.* |
| Ethics oversight | *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.* |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

| | |
|---|---|
| Clinical trial registration | *Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.* |
| Study protocol | *Note where the full trial protocol can be accessed OR if not available, explain why.* |
| Data collection | *Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.* |
| Outcomes | *Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.* |

## Dual use research of concern

Policy information about dual use research of concern

### Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

| No | Yes | |
|---|---|---|
| ☐ | ☐ | Public health |
| ☐ | ☐ | National security |
| ☐ | ☐ | Crops and/or livestock |
| ☐ | ☐ | Ecosystems |
| ☐ | ☐ | Any other significant area |

### Experiments of concern

Does the work involve any of these experiments of concern:

| No | Yes | |
|---|---|---|
| ☐ | ☐ | Demonstrate how to render a vaccine ineffective |
| ☐ | ☐ | Confer resistance to therapeutically useful antibiotics or antiviral agents |
| ☐ | ☐ | Enhance the virulence of a pathogen or render a nonpathogen virulent |
| ☐ | ☐ | Increase transmissibility of a pathogen |
| ☐ | ☐ | Alter the host range of a pathogen |
| ☐ | ☐ | Enable evasion of diagnostic/detection modalities |
| ☐ | ☐ | Enable the weaponization of a biological agent or toxin |
| ☐ | ☐ | Any other potentially harmful combination of experiments and agents |

## ChIP-seq

### Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

| | |
|---|---|
| Data access links<br>*May remain private before publication.* | *For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.* |
| Files in database submission | *Provide a list of all files available in the database submission.* |
| Genome browser session<br>(e.g. UCSC) | *Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.* |

### Methodology

| | |
|---|---|
| Replicates | *Describe the experimental replicates, specifying number, type and replicate agreement.* |
| Sequencing depth | *Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.* |
| Antibodies | *Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.* |
| Peak calling parameters | *Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.* |
| Data quality | *Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.* |
| Software | *Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.* |

# Flow Cytometry

## Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

## Methodology

| | |
|---|---|
| Sample preparation | *Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.* |
| Instrument | *Identify the instrument used for data collection, specifying make and model number.* |
| Software | *Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.* |
| Cell population abundance | *Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.* |
| Gating strategy | *Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.* |

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

# Magnetic resonance imaging

## Experimental design

| | |
|---|---|
| Design type | *Indicate task or resting state; event-related or block design.* |
| Design specifications | *Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.* |
| Behavioral performance measures | *State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).* |

## Acquisition

| | |
|---|---|
| Imaging type(s) | *Specify: functional, structural, diffusion, perfusion.* |
| Field strength | *Specify in Tesla* |
| Sequence & imaging parameters | *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.* |
| Area of acquisition | *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.* |

Diffusion MRI ☐ Used ☐ Not used

## Preprocessing

| | |
|---|---|
| Preprocessing software | *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).* |
| Normalization | *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.* |
| Normalization template | *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.* |
| Noise and artifact removal | *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).* |

Volume censoring

*Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

## Statistical modeling & inference

Model type and settings

*Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested

*Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)

*Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction

*Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

## Models & analysis

| n/a | Involved in the study |
|---|---|
| ☐ | ☐ Functional and/or effective connectivity |
| ☐ | ☐ Graph analysis |
| ☐ | ☐ Multivariate modeling or predictive analysis |

Functional and/or effective connectivity

*Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

Graph analysis

*Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

Multivariate modeling and predictive analysis

*Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*